

Conflict of interest revisited

As the number of academic biologists with ties to business increases, the pool of independent scientists for service on advisory boards and peer review panels is dwindling. Will it eventually disappear?

THE biomedical enterprise in the United States is uniquely dependent on the independence of its academic scientists. The National Institutes of Health (NIH) depend on university scientists for the peer review panels (known as study sections) which sit in judgement on the thousands of grant applications considered each year. The National Academy of Sciences (NAS) depends on the expert opinion of academic researchers who serve on its study boards and panels. The Food and Drug Administration needs the advice of independent counsellors for decisions ranging from the safety of silicone breast implants to the effectiveness of new classes of drugs such as cytokines.

Although review panels of various kinds often include representatives from industry or special interest groups, academic scientists have been the backbone of the massive advisory system in the United States, as they have elsewhere. Whether that will continue to be true is in doubt as more and more prominent biomedical scientists forge ties with the business world. Two decades ago, before the recombinant DNA revolution, the good guys were not in science for the money.

This laudable position was the norm as long as the opportunities for biomedical scientists to make money in industry were few and far between. It is easy to resist a temptation that is not there. It is by now old news that the rapid growth of the biotechnology industry has changed all that. As we note in this week's News pages (page 753), the NAS has had to establish a new advisory category (called *liaison*) for people whose expert advice carries with it the burden of a financial conflict of interest.

The conflict-of-interest debate, which requires urgent resolution if the public and Congress are not to lose confidence in academic science, is going to be quite difficult to resolve because two equally important social values are put in opposition to each other. On the one hand is the familiar expectation that scientists who offer advice on matters of public policy do so without an eye to personal gain. The other imperative is that the fruits of basic research, supported as it is by the taxpayers' money, should be promptly transferred to the marketplace and the bedside.

The importance of technology transfer to the public interest and the US economy (under the buzzword of international competitiveness) was not long ago spelled out in law. The little-read but nonetheless significant Technology Transfer Act requires federal scientists to seek patents on applicable discoveries and encouraged

researchers at NIH and other government laboratories to find ways of collaborating with colleagues in private companies. Thus, at the NIH, there is a proliferation of CRADAs (cooperative research and development agreement) that bring industry scientists to NIH laboratories and vice versa. The rules governing the ethical boundaries of CRADAs (from which NIH scientists make no personal money) are so elaborate that, sooner or later, someone is going to be in violation and a scandal will erupt.

What is the solution? There is no chance of going back to the days when biologists (unlike their colleagues in physics and engineering) lived in an ivory tower. There has not been much fuss in the hard sciences over conflict of interest resulting from collaborating with industry, but that may be attributed to history rather than fundamental differences in the relationships. So, is it likely that conflict-of-interest can be accepted in the biomedical enterprise? No. Even though many scientists now have ties to the profit-making sector, few are entirely comfortable with their new status. The culture has not yet changed entirely. Would full disclosure of business interests do the trick? Not altogether if the NAS's 'liaison' category is any indication.

Perhaps the nature of advisory bodies will have to change. It is common to ask the most prominent (and senior) members of the community to serve. Will boards now seek members from the ranks of younger scientists? Will the market eventually provide a solution? If there is a scarcity of able people free to serve on public boards, those that survive may have to be paid for their services, perhaps handsomely. And boards may scrape by with fewer members (not necessarily disastrously). But the long-run problem is a problem not for those who serve on public boards, but for the academic institutions from which they come. The universities' duty to the prosperity of the community in which they are embedded, may seem to have shifted to the enrichment of their researchers. Can that go on?

Ever since academic biologists took their first tentative steps toward industry in the late 1970s and early 1980s, when companies such as Hoechst, DuPont, and Monsanto were contributing millions of dollars to university collaborations, the research community has worried about its academic soul. There is no easy way out of the present dilemma, which will only get worse during coming years. It is an issue that could well tear the community apart. □



Saving the world

English-speaking academies have stated a global problem, but have yet to answer it.

"It's not what we say, but who we are." That, legitimately enough, is the unspoken message behind last week's joint statement of the Royal Society and the US National Academy of Sciences on population growth and other global problems (see page 759). Although there is little in the statement that has not been said before, the prestige of both academies is bound to carry substantial weight. The fact that both institutions have been silent through the often clamorous debates of the past quarter of a century can only increase the chances that they will now be listened to with care. Many who have hitherto shut their eyes to global problems will now be drawn in. So far, so good.

Exactly similar considerations have no doubt caused the academies to bite off their tongues at several points in their statement at which readers would expect them to have urged specific remedies. It is easier (although not easy) to win agreement among the diverse memberships of academies on the statement of important problems than to secure assent to specific recipes for their solution.

Thus, few will be offended that the academies now urge the development of new generations of "contraceptive agents and devices", but there might well have been ructions if the academies had complained of the US government's refusal to finance aid programmes that support abortion, even indirectly, not to mention the policy of churches such as the Roman Catholic designed to inhibit the use of contraceptives. And while there will be wide liberal endorsement for the academies' implicit view that the best way of managing population growth in developing countries is to effect rapid demographic transition, the serious impediments are the perverse obscurantism of many of the governments concerned and the enormous sums of money that industrialized countries would otherwise be required to transfer to them. Must it always be for others than academies to make those arguments?

The issue of priorities is also clouded. To be sure, the statement is right to single out the threat of global warming for attention. (If Mr John Sununu were still at the White House, Dr Frank Press at the National Academy would have a sharp note in the morning.) But the cause of maximal biodiversity does just as well as global warming, but unjustifiably. There are, of course, all kinds of reasons why species should be conserved. Some are ecological, and direct determinants of human well-being; it would, for example, be a thoroughly bad business if plants of the species that make cereal crops were put in hazard. In other cases, as in the conservation of whales and tigers, sitting at the tops of independent food chains, the case for conservation is, rather, aesthetic. There is also a strong thread of opinion, not widely shared, that all species have an equal right to careful custodianship. The better appraisal of the extent of global biodiversity for which the academies ask will make it easier to tell which species matter most. Would

the academies then agree that sheep, say, might be distinguished from, say, goats?

The position of biodiversity reflects a more general difficulty with the academies' general statement of their global problem. To the extent that solutions must rest on international agreements, explicit (as in the Montreal Convention) or implicit (as when donor governments let recipients know what kind of government they expect of them), the best that can be done is the most of what is feasible. One issue is whether the conference at Rio de Janeiro in June should aim at a treaty of greenhouse gas emissions in which specific limits are included, or whether instead it should aim at including as many states as possible, with the understanding that specific limits should come later (and when they can be enforced). These are essential political, not technical, decisions. Who, for example, could expect the states of the former Soviet Union to sign a global warming treaty with specific limits just a few months from now? And who would believe them if they did? This does not imply that the ideal is unattainable, but merely that it may take a little longer.

There are still more divisive issues that deserve attention, not least at the conference the academies plan to organize next year. In the past few years, it has become commonplace for financial aid to developing countries to be provided only for projects that are environmentally acceptable. The intention is laudable enough, but it supposes that all environmental damage is irreversible and that it is an acceptable charge on the meagre funds available for assistance in developing countries that a substantial fraction should be hived off for environmental purposes. The general application of this policy by agencies such as the World Bank would be mistaken. Again, the question should be that of the goals to which attention should most urgently be paid. If the objective is to reduce fertility, reductions of infant and childhood mortality are what matter more than anything else.

In the circumstances, it is honest but overdefensive of the academies to proclaim that the world should not look to science and technology alone for a solution of all these problems. The truth is that the problems to which the academies have set their pens are economic, social and political, with ethical issues never far beneath the surface. For what it is worth, much the same was true in the heyday of nineteenth century technology, and the beginning of most of the academies' global problem. Who would have been able to afford James Watt's steam engines if the banking system had not already been invented? How would the breakneck exploitation of innovation in this century have been possible without the invention, in developed countries, of social security systems? The truth is that, even in untrammelled growth, technology was not autonomous. Why should it be when the restraints are on? □

Correction

Mr Uri Geller asks that the erroneous reference to a Florida court in a recent leading article *Nature* 355,284; 23 January 1992) should be corrected. The trial of his libel action against Mr James Randi is before a federal court in Washington DC. The impending proceedings are not the trial itself, but part of the deposition (disclosure) process. □

Conflict concerns disrupt panels, cloud testimony

- Researchers resign from Academy committees
- DNA fingerprinting suit alleges undisclosed ties

Washington

DURING the past year and a half, six members of two genetics panels at the National Academy of Sciences (NAS) have resigned or been asked to drop their connections with private companies because of concerns that commercial ties might affect their scientific judgement. This unprecedented rash of concern about conflict of interest sharpens the debate over how to separate science and commerce. The relationship is especially murky within genetics, where today's scientific discovery may be tomorrow's commercial products.

While many in the scientific community continue to play down the problem of conflict of interest, the NAS moves have not gone unnoticed. In Congress, Representative Don Edwards (Democrat, California), chairman of the House Judiciary Committee's subcommittee on civil and constitutional rights, plans some time this spring to hold hearings over concerns about conflict of interest among the scientific proponents of DNA 'fingerprinting'. Earlier this month, lawyers who argued (and lost) a landmark DNA fingerprinting case in 1989 petitioned to have the case reopened, saying that geneticists who gave expert testimony did not disclose their

commercial ties. And the Council for Responsible Genetics, a research watchdog group, wants tougher ethical standards.

At the NAS, the problem of conflict of interest among panel members has become so serious that officials have been forced to create a new class of committee — a nonvoting 'liaison panel' — exclusively for affected researchers. When the NAS's Institute of Medicine (IOM) reviewed the membership of its new genetic screening panel (known as the Committee on Predicting Future Diseases) last year, it asked the chairman, C. Thomas Caskey, and two other panel members to step down because of concern about their ties to industry. The two panel members, Frank Fujimura and Philip Reilly, agreed to sit on an associated liaison panel, but Caskey decided to leave the committee entirely.

Even when the IOM picked a new chairman for the genetic screening panel, conflict of interest once again reared its head. Before Arno Motulsky, a University of Washington geneticist, could take the post, IOM officials asked him to resign from the advisory board of GeneScreen Inc., a major genetic screening company, and sell his stocks in the company. He says he agreed out of a "sense of duty to IOM"

and because the amount of money involved was relatively small.

But the debate over conflict of interest in genetic testing pales compared with the open warfare now being waged over DNA fingerprinting, where ideology (law and order vs civil rights) and the prospects of substantial profits have splintered the genetics community. At the centre of the controversy is Caskey, a Baylor College of Medicine geneticist and a tireless promoter of DNA fingerprinting as a forensic technique. On 21 December 1991, six months after he had resigned from the IOM genetic screening panel and two days after an article in *Nature* (354, 500, 1991) described his financial connections to a major DNA fingerprinting company, Caskey resigned from a NAS panel doing a report on DNA fingerprinting.

On 8 February 1992, the debate over Caskey's conflicts reached the courts when defence lawyers in a 1989 murder case known as *US v. Yee, et al.* filed a motion for the case to be reopened. Their motion was based on what they allege was a conflict of interest in expert witnesses, including Caskey, who testified in support of the DNA fingerprinting analysis that sent three men to prison.

Attorneys Barry Scheck and Peter Neufeld allege that Caskey failed to disclose that he had already applied for a grant for \$200,000 from the National Institute of Justice (NIJ), which is funded by the US Department of Justice, to do DNA fingerprinting research. Although Caskey declined to be interviewed, he submitted an affidavit in the *Yee* case in which he attests that he did not "personally profit"

Coincidence or conspiracy?

In their wide-ranging attack on DNA fingerprinting and its supporters, the two lawyers who have asked to have a landmark DNA fingerprinting case reopened (see story above) are alleging that the *American Journal of Human Genetics* (AJHG) allowed Federal law enforcement officials to interfere in the peer-review process of a critical paper.

In the paper, Seymour Geisser, a University of Minnesota statistician, criticizes the statistical methods used by the Federal Bureau of Investigation (FBI) in DNA analysis. Late last year Geisser submitted his manuscript for publication in the *AJHG*. On 15 January, Geisser received a fax from Stephen Redding, a federal prosecutor involved in a Minnesota DNA fingerprinting case in which Geisser was expected to testify for the defence. Redding demanded that Geisser bring to court any of his papers on DNA fingerprinting now pending in any peer-reviewed journal, specifically including the *AJHG*, and all correspondence concerning the paper.

Fifteen minutes later, Charles Epstein, editor of the *AJHG*, faxed Geisser a covering letter warning him to write carefully as the work would certainly be used in court cases, and attached three peer reviews, one of which was exceedingly critical. Late last month, in a court hearing for another DNA fingerprinting case, Ranajit Chakraborty, a population geneticist at the University of Texas Health Science Center, was forced to identify himself as the critical reviewer of the Geisser manuscript. At the hearing, Chakraborty also acknowledged that he

is a co-investigator on a \$200,000 Justice Department grant that is principally intended to generate a series of peer-review articles that would provide better scientific support for the FBI's statistical methods.

The *Yee* lawyers allege that the 15-minute proximity of the two faxes was no coincidence. They allege that at least one of the reviewers tipped off federal prosecutors, and that the prosecutors essentially ran the peer-review show from there.

Nonsense, says Epstein. He says that "nobody contacted us and we contacted nobody." The letter and reviews were faxed to Geisser two days after the third review came in, a time span entirely determined by Epstein's own schedule, he says.

More worrying, however, is the fact that Epstein says he did not know until it was revealed in the court documents earlier this month that Chakraborty, who is an *AJHG* associate editor, had the Justice Department grant. Had he known, he says, "I'm not sure I would have used him as a reviewer."

Bias and conflicts in this field of inflamed passions are probably unavoidable; they become a problem only when researchers do not disclose them. With hidden agendas and undisclosed allegiances, Epstein says, "it is becoming increasingly difficult to deal with these papers." The only reason Chakraborty's Justice Department connections came out was because a court required him to disclose them. Just how many potential conflicts in less controversial fields lie hidden remains a nagging question.

Christopher Anderson

from the NIH award, which, like most grants, was awarded to his institution.

The defence motion in *Yee* also discloses that another expert witness, Stephen Daiger of the University of Texas Health Science Center, failed to disclose that he had applied for and subsequently won a \$300,000 grant from NIH for DNA forensics research. Ranajit Chakraborty, co-investigator on the NIH grant with Daiger, says that one of the principal purposes of the grant was to generate a series of peer review articles that would substantiate the FBI's statistical methods.

In a response on 20 February to the petition to reopen the case, James Wooley, assistant US attorney for the US district court in Ohio, rejected the motion as "an all-out bare knuckles, take-no-prisoners assault" on DNA fingerprinting by the "anti-DNA lobby". Those forces, he claims, are trying to "pollute the record with allegations". Because opponents of DNA fingerprinting have usually failed to have DNA evidence dismissed on scientific grounds, he says, they have decided to attack its proponents.

Although the DNA fingerprinting fracas is perhaps the most high-profile of the recent conflict of interest debates, some of the other cases at the NAS have been just as vexing for the researchers involved. Last summer, NAS officials decided to take a closer look at the members of the new genetic screening panel.

After reviewing the standard "conflict and bias" declarations submitted by panel members, NAS officials decided that Reilly, who is executive director of the Shriver Center for Mental Retardation in Waltham, Massachusetts, and Fujimura, scientific director of molecular biology at the Nichols Institute Reference Laboratory in San Juan Capistrano, California, each had a potential conflict of interest. Reilly serves as a member of the board of directors of Vivigen Inc., a diagnostic testing laboratory in Santa Fe, New Mexico, and Fujimura's company is a commercial laboratory that does genetic screening and DNA fingerprinting.

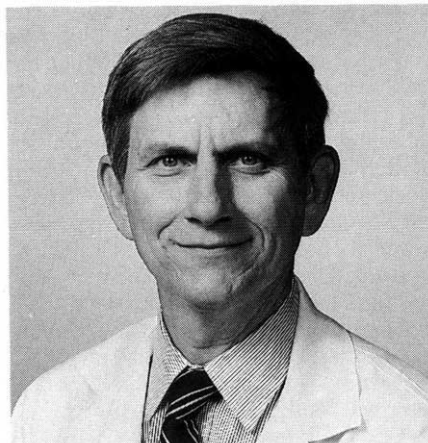
Reilly, for one, was surprised. Neither he nor Fujimura has ever made a secret of his commercial affiliations, he says, and the NAS DNA fingerprinting panel, on which he also serves, had not raised any questions about conflict. In fact, of the eight national advisory panels on which he serves (including those of the National Institutes of Health and the Office of Technology Assessment), only four even required a conflict-of-interest review, he says. And only one institution — the IOM — decided that there was a conflict.

Although Reilly agrees with the IOM decision, he's troubled by the random nature of most reviews of potential conflicts of interest. What is a conflict for one advisory committee may not be one for another, even when they cover the same

subject. "The problem is the unevenness of it all," he says.

In some of the other cases, the initial concern came from the researchers themselves. Michael Hunkapiller resigned in 1990 from the NAS DNA fingerprinting panel when he decided that his job as vice president for science and technology for Applied Biosystems Inc. (ABI), a genetics technology company, had become a potential conflict.

He first thought that his employment at a company that made only "generic technology" in the field would not pose a problem in weighing DNA fingerprinting as a member of the NAS committee. But



Caskey's industry ties are under fire.

when ABI decided to market a machine that was specifically applicable to DNA fingerprinting, Hunkapiller asked to be taken off the panel.

But Caskey's case is perhaps the most difficult of all for a scientist. He is both at the forefront of basic genetic research, and deeply involved in the growing genetic testing and diagnostics industry. Beyond the NIH grant, he has an agreement to license genetics diagnostics technology through Baylor to Cellmark Diagnostics Inc., a major DNA fingerprinting company that was asked by the US Department of Defense to develop methods to identify soldiers killed in the Gulf War. Two other companies are about to conclude similar licence agreements. Royalties from these agreements go to Baylor, not to Caskey directly, according to a spokeswoman.

Caskey is also connected to industry through his wife, Peggy Caskey, who in 1982 founded (and now serves as president of) the Houston-based Laboratory for Genetic Services, Inc., a commercial diagnostics laboratory. Victor McKusick, a Johns Hopkins University geneticist and chairman of the NAS DNA fingerprinting, says this connection was not the panel's principal concern.

"I don't think that what a member of your family does in this regard can be regarded as a conflict," he says. As for the other potential conflicts, McKusick says,

"it may be a matter of an oversight here. I think that Caskey's a honest and honorable man." He describes Caskey's affiliation with Cellmark during the Gulf War as a "patriotic duty".

One congressional staff member who has been following the controversy sees it differently. "I don't understand why it wasn't apparent that Caskey had a conflict," he says. "The academy obviously lost control of their process." The House Science, Space and Technology committee had a meeting with academy officials to discuss the issue of conflict and bias in academy reports but, unlike Edwards's justice subcommittee, does not currently plan hearings on the subject.

Although the incidence of conflict appears to be growing, NAS officials say that there is little they can do to uncover potential problems beyond the standard precaution of having each panel member fill out a self-disclosure form, which is updated yearly. They can, however, make sure that the disclosure forms do not become just another formality.

Oskar Zaborsky, IOM staff officer for the DNA fingerprinting panel, asked the panel members to reexamine their potential conflict in December, after the Caskey controversy. And Ruth Bulger, who heads the IOM section in which the genetics screen panel serves, says that she would have "asked earlier" in the process about possible conflicts if she had known that three of the members would be found to have potential problems.

The Council for Responsible Genetics, based in Cambridge, Massachusetts, is leading the internal battle against conflict of interest in the genetics community. As part of a letter-writing campaign to researchers and journals, the council warns that the Human Genome Project is at risk of becoming "compromised by the undisclosed biases of its participants." Paul Billings, a geneticist at the California Pacific Medical Center and member of the Council, says that as a reviewer for the \$5 million ELSI programme, he has seen several instances of researchers submitting grant proposals that included payments to private companies (such as commercial testing laboratories) in which they had undisclosed financial interests. "It's ironic that the world's largest ethic programme would have this problem," he says.

Although the problem of conflict has been nowhere as apparent as in DNA fingerprinting, where critics and supporters trade allegations of bias and dirty politics, researchers are predicting more battles in genetic screening. As the Human Genome Project begins to fulfil its promise, more geneticists will have to decide just where to draw the line between commerce and pure research — and groups such as the NAS will have an even harder time finding impartial advice.

Christopher Anderson

Panelist cries foul

Washington

WHILE the National Academy of Sciences and the courts are struggling with financial conflict of interest, the US Food and Drug Administration (FDA) has run into a case of plain old outspokenness.

Earlier this month, FDA deputy commissioner Carol Scheman asked Norman Anderson, a Johns Hopkins University physician, to give up his voting status on a FDA advisory panel on breast implant safety. She says it was because *Time* magazine published excerpts from a letter to the FDA, dated 12 December, in which Anderson warned of a "breach of trust" by the implant's manufacturer, Dow Corning Wright. Those words suggested to Scheman that he had "already reached a conclusion" about the safety of the implants before the panel heard all the evidence.

That is probably true. Anderson has long been a critic of breast implants, and his letter of 12 December was written after he had reviewed internal company documents that suggested a cover-up by Dow Corning Wright. But as a past chairman of two previous FDA panels on breast implants, he is also well informed. And he has never made a secret of his opinions.

His punishment has drawn fire from such critics as Representative Ted Weiss (Democrat, New York), who claims that FDA is running dangerously close to infringement of the First Amendment right to free speech. In past months, FDA has similarly come under criticism for attempting to restrict the kind of scientific conferences that drug companies can sponsor and for censoring press releases and video advertisements from the pharmaceutical industry (see *Nature* 354, 421; 1991).

Scheman says that in this case Anderson appeared to have a "predisposition" against breast implants, which for the purposes of the panel was of just as much concern as a conflict of interest.

However, in a letter dated 13 February to FDA commissioner David Kessler, Anderson says that there is more to the FDA position than Scheman is admitting. In particular, he says that Scheman called him and asked him to relinquish his voting status on the panel a week before the *Time* article came out. According to Anderson, Scheman said that her definition of "conflict of interest includ[es] 'knowing too much' " about subjects under review.

In a letter of 13 February to Weiss, who has taken up Anderson's case, Scheman attempted to clarify the FDA's position: "Access to information does not create a problem...unless the member uses that information either to reach a conclusion concerning the subject of the Panel's review...or has made public statements which create the appearance that such a conclusion has been reached."

Christopher Anderson

'Big three' test the waters

London & Tokyo

A LITTLE-NOTED but potentially huge milestone in the history of research collaboration between the United States, Europe and Japan was reached this week when government officials, industrialists and academics gathered on 24 February in Toronto to launch a two-year pilot study on intelligent manufacturing systems (IMS). This is the first major collaborative programme of commercial significance involving the world's three main scientific blocs.

The IMS project has had a difficult and prolonged birth. Proposed two years ago by the Japanese Ministry of International Trade and Industry (MITI) (see *Nature* 343, 496; 1990), IMS is intended to apply cutting-edge information technology to industrial manufacturing. It was designed as a counterpart to the International Human Frontier Science Program, a MITI initiative that is now jointly supported by the G-7 nations and the European Communities and that supports basic research in molecular biology and neuroscience. Both programs are Japanese attempts to put money into collaborative research and to counter criticism by Western countries that Japan's economic success has been achieved by 'riding piggy-back' on the research expertise of Europe and the United States.

Because Western governments initially feared that the Frontier program was an attempt to pick the brains of their researchers, MITI officials decided that their next initiative would cover an area in which Japan is already acknowledged as a world leader. But the IMS project also became bogged down amid suspicions about Japan's motives, and the basic formula for collaboration worked out for the pilot project starting this week differs markedly from MITI's original plan.

MITI envisaged having \$1,000 million to spend over 10 years. Japan was to provide 60 per cent of this money, with the rest divided between the United States and the European Communities (EC). Most of the work would have been carried out in a single new research centre, set up in the EC or the United States, although the programme would have been administered from Tokyo.

But now the proposed international fund, the central administration and the research centre have all gone. Instead, officials were expected this week to choose three pilot collaborative research projects to be conducted in the home laboratories of the researchers involved. The nations taking part (the EC, United States and Japan have now been joined by Australia, Canada and the countries of the European Free Trade Association) will finance the work carried out on their own territory.

EC officials say that MITI's original plan seemed to offer Japan a competitive advantage. A central office in Tokyo would have given the Japanese an "immense wealth of information". The current plan, they say, should ensure that the three main partners benefit equally from the project.

But some observers believe that the US government is less pleased with the new approach. The negotiations have progressed "with Japan with its foot on the gas and the United States with its foot on the brake," according to one European Commission official. US officials are said to have opposed the central administrative office in Tokyo, but favoured the notion of Japan footing a hefty proportion of the bill.

Phyllis Genther, director of the US Department of Commerce's Japan Technology Program, offers a different explanation. She says that US negotiators wanted each of the partners to contribute equally; in particular, they asked that each should bring a similar package of new technology to the project, and not just money.

For their part, the Japanese proponents of IMS say they can live with the revised plans. Hidehiko Nishiyama, deputy director of MITI's industrial machinery division, says that Japan has plenty of experience with "decentralized" programmes through MITI's own domestic research schemes.

The big question now is whether the IMS formula will serve as a model for further collaborative ventures between Japan, Europe and the United States. A lot of people "who don't give a damn about IMS" will be keeping a close eye on the two-year feasibility study, says one European Commission official. But both the United States and the EC are withholding judgement on the value of the IMS model for future collaboration until the details have been more fully worked out.

Peter Aldhous & David Swinbanks

FAISAL PRIZE

Two honoured

London

Two European scientists have won the 1992 King Faisal International Prize in science and in medicine.

Sydney Brenner, director of the UK Medical Research Council's Molecular Genetics Unit in Cambridge, was awarded the science prize for his work in discovering the triplet codon genetic code, and demonstrating the existence of messenger RNA. This year's medicine prize has gone to Attilio Maseri, from the Catholic University of Rome, for his work on coronary artery disease. Brenner and Maseri will both receive \$93,000.

Peter Aldhous

US picks 2 sites, Europe waits

Washington

THE United States has selected two sites for its \$211 million gravity wave detector, a step that European scientists hope will encourage their governments to move ahead with plans for similar facilities.

Last week the National Science Foundation (NSF) announced that it has chosen Hanford, Washington, and Livingston, Louisiana, as twin sites for the Laser Interferometer Gravitational Wave Observatory (LIGO). The reality of gravity waves, essential consequences of Einstein's general theory of relativity, has so far not been directly demonstrated.

The US project, directed by Rochus (Robbie) Vogt of the California Institute of Technology, will consist of two detectors, each with twin 2.5-mile arms arranged at a 90-degree angle. The sites are in rural and seismically stable areas, where development is unlikely, to reduce the chances that external vibrations will disturb the delicate instruments.

The Washington site, in the south-central part of the state, and the Louisiana site, 30 miles east of Baton Rouge, were judged the best of some two dozen potential sites scattered across the country. NSF director Walter Massey said that his criteria included accessibility, ease of land acquisition, and the promise of strong ties to local universities. The Hanford site is part of a vast government tract where nuclear fuel was first produced and where waste is now stored, while the Louisiana site is part of a private tract of forest, now being harvested for paper products, that will be deeded over to the government.

The United States is spending \$23 million this year to begin work on the two sites. The administration has asked Congress for \$48 million in fiscal year 1993 (which begins on 1 October 1992) to continue construction, which is expected to take five years. Vogt says the two sites will be built back-to-back so that the same team can be used to install the instrument, which at its core is a weight, hung at the end of a four-foot wide vacuum tubes, that is exposed to a continuous beam of laser light.

Vogt says that the US project can proceed without any help from the Europeans but that eventually a third and possibly a fourth site will be needed to perform the triangulation needed to pinpoint the celestial location of any waves that are detected. And he is optimistic that progress here will help his colleagues in the four countries — France, Italy, Germany and Britain — that have been asked to contribute to similar observatories in Europe.

"They are slaves to decisions in the US," says Vogt. "They are behind us now, but I think that they can catch up quickly once they get started."

The European projects have travelled

along a rocky road since they were first proposed. The original plans were for a French-Italian collaboration, called Virgo, to be built near Pisa in northern Italy, and a German-British collaboration, called Geo, to be built near Hanover. But with six agencies — three in Germany alone — being asked to contribute funds, progress has been slow.

Last fall Britain's Science and Engineering Research Council (SERC) took a decidedly hostile view of its proposed \$10 million contribution to the \$90 million project. Since then, physicists from the four countries have formed a more visible common front, called 'Eurograv', to promote the idea of a genuine European collaboration.

In the meantime, Virgo is moving ahead, says team leader Allain Brillet of the University of Paris at Orsay. Next month the group will present its final design proposal to a joint committee of the two funding agencies. A decision from them is expected in June, and funds for the project could be approved in the autumn

for release in January. Nearly 100 physicists and engineers are already at work on the project.

The outlook for the Geo project is somewhat less certain, admits team leader Karsten Danzmann of the Max Planck Institute for Quantum Optics in Garching. Last week a scientific panel on proposed large research projects submitted a report to Germany's research minister, Heinz Riesenhuber, that spoke favourably of the gravity wave detector but did not place it as one of the country's highest research priorities. Danzmann admits that German unification has put a heavy strain on his country's budget, and that Germany's existing commitments to other joint European projects, as with the proposed Large Hadron Collider at CERN, may require substantial new funding.

Danzmann nevertheless thinks that a European gravity wave detector is inevitable. "I'm convinced that money will be forthcoming in the next couple of years," he says. "Last year I would have said we'd be ready to start in a few months. Now it looks like it might be 1995 before everything falls into place."

Jeffrey Mervis

RUSSIAN SCIENCE

A way to wean weapons makers

Washington

OFFICIALS from the United States, Germany and Russia meet today (27 February) in Bonn to flesh out last week's announcement that the three countries will create an international centre to support civilian research by Russian nuclear weapons scientists and engineers. There are as yet few details about the centre's structure, its methods of soliciting and selecting grants, the pool of eligible scientists or the type of work to be supported.

US government research agencies, including those from the National Science Foundation and the Energy Department, admit that so far they have played almost no role in the discussions. The first news of the centre came at a press conference on 17 February that followed a three-hour meeting in Moscow between US Secretary of State James Baker and Russian president Boris Yeltsin. Later that day their governments issued a joint statement with the German Foreign Minister, Hans-Dietrich Genscher, who last month suggested some type of structure to support the estimated 2,000 or so scientists who occupy the top layer of the nuclear weapons complex of the former Soviet Union.

"It's not too early to ask these questions, but it's much too early to provide any answers," says a spokesman for the German embassy in Washington. "The first brainstorming will take place in Bonn."

This week's meeting provided the major donor countries with their first chance

to discuss the proposed centre. It will be followed by a larger meeting next month of all participating countries, and subsequent meetings as needed.

The United States has pledged \$25 million towards the centre, which will be staffed by science administrators from both donor countries and the Commonwealth of Independent States. At least twice that much is needed to operate the centre, however, with contributions expected from industry and private foundations as well as from national governments.

The new centre is intended to put the former Soviet scientists to work on a broad range of civilian projects covering both basic and applied research. "The whole point of this is to fund projects that have nothing to do with nuclear weapons," says a State Department official. Baker was bombarded with ideas during a recent visit to a formerly closed Soviet weapons facility, and scientists from the participating countries will also be encouraged to submit proposals for joint efforts with their Russian colleagues.

The centre is more likely to support proposals of limited scope and short duration than larger and more costly efforts with a timescale of five or ten years. "We want to support as many researchers as possible," says the State Department official, "as soon as possible." If all goes well, the centre could be up and running by the summer.

Jeffrey Mervis

Parties discover technology

London

WITH the whiff of a general election in the air, the two main British political parties have rushed to endorse a new recipe to boost technology transfer from academia to industry. In fact, their rapid response to what is only a draft report is a bit worrying to its author, Sir John Fairclough, formerly chief scientific adviser to the UK government. And it could become a divisive issue for those research organizations that would be most affected if its suggestions were carried out.



John Fairclough: worried

Reports on technology policy do not customarily receive such an instant political reaction. But then again, they are not usually released in the middle of a (still unofficial) general election campaign, when the political parties are keen to take on board any potentially vote-winning initiative.

"It's rather nice to have both political parties supporting what you're doing," says Fairclough, chairman of the working party on innovation that wrote the report. "I just hope that it doesn't get screwed up."

The report argues that Britain needs a network of research institutes, doing contract work for industry, but closely associated with a university or polytechnic. These proposed 'Faraday Centres' are modelled loosely on the 35 German Fraunhofer Institutes — applied research centres with a total staff of almost 6,000, including some 1,200 graduate students.

Fairclough's group, including many of Britain's leading science policy experts, was asked last year by the Prince of Wales to look at ways to bridge the gulf between British companies and universities. That gap is thought to be an important factor behind the common observation that many valuable technologies were invented in Britain, but exploited elsewhere.

Within hours of the report's release last week, the opposition Labour party issued its own technology plan, including an almost carbon-copy proposal for a series of 'Newton Institutes'. Labour science and technology spokesman Jeremy Bray says that his party's technology proposals are the result of two years of thought and consultation. But Bray conceded that the Newton Institute proposal is "remarkably similar" to the plan put forward by Fairclough's group.

The Conservative government, with the public purse at its disposal, was able to go one step further. The industry secretary, Peter Lilley, and Kenneth Clarke, his counterpart at the Department of Education and Science, announced a £2 million pilot programme to encourage five existing research organizations to take on up to 20 graduate students each in the coming year.

The document released by Fairclough's group last week was merely an interim report, and contains little detail about the number of graduate students who should be supported in the Faraday Centres, or how the programme would be funded. Fairclough says he would have preferred the government to wait until these details had been fully worked out before moving ahead with a pilot programme. The Centre for Exploitation of Science and Technology (CEST), a government- and industry-funded think-tank chaired by Fairclough, is now working on a more detailed plan.

Arguing that technology transfer is best achieved by the movement of people, Fairclough's group wants the best of Britain's existing independent research organizations to take on graduate students. The idea is that they would work towards their doctorates on industrially relevant research. Such training, so the theory goes, would make them more likely to move into industry than students in a traditional academic environment.

Last week's events may also force a change of plan at the Science and Engineering Research Council (SERC). The council, which funds more British graduate students than any other body, can expect to become closely involved with any Faraday Centre or Newton Institute initiative. SERC is working on a new graduate training scheme, leading to a doctorate, that is aimed specifically at young engineers.

The Fairclough group says that SERC's plan is complementary to its proposals. But SERC officials say they had planned to award these grants in the universities or polytechnics, and have not been consulted about the Faraday Centres idea.

There are already several dozen institutions in Britain that fit the mould of what is called for in the Fairclough report. John Bennett, secretary-general of the Association of Independent Research and Technology Organizations (AIRTO), whose 36 members mostly work on specialized contract research projects for industrial clients, says he is pleased that both political parties now recognize the need to strengthen the tier of research institutions between the universities and industry. "[Former prime minister Margaret] Thatcher believed you can bend academia and industry and get them to meet in the

middle," he says. But that approach, Bennett believes, has not succeeded.

Nevertheless, the Faraday/Newton proposals would require companies that belong to AIRTO to become much more involved in training graduate students. Bob Whelan, chief executive of CEST, believes that up to 25 per cent of the staff of a Faraday Centre could be graduate students. But Bennett estimates that only about 50 of the 10,000 people that AIRTO members employ fit that description.

Many AIRTO members may also be reluctant to take on a large number of graduate students. Their fear is that academic influence may blunt their commercial edge. The Faraday/Newton proposals also imply the existence of a government-linked umbrella body to co-ordinate the scheme, which may put off some of the more fiercely independent of AIRTO's members.

Peter Aldhous

EVOLUTION EDUCATION

Creationist victory

CALIFORNIA state education authorities last month agreed to pay \$225,000 to a creationist graduate school and to end any attempt to strip the school of its authority to grant masters degrees in science. In an out-of-court settlement, state lawyers agreed that the Institute for Creation Research (ICR) graduate school and other private, post-secondary educational institutions "may teach the creation model as being correct, provided the institution also teaches evolution."

"It's a shocking and puzzling decision", says Kevin Padian, a biologist at the University of California, Berkeley who lobbied against the ICR. "The state lawyers gave away the store." California is revising its education policies, he says, and state officials were unwilling to do battle with religious groups during this transition period.

Although the ICR legal challenge was based mostly on technicalities, the state lawyers agreed to a list of concessions that, among other things, would prevent Bill Honig, the state education superintendent, from participating in decisions about future ICR licences. Honig has ordered two scientific reviews that led to a recommendation to strip ICR of its degree-granting authority (see *Nature* 343, 501; 1990). Last month's settlement deletes those reports from the record.

Kenneth Cumming, dean of the ICR graduate school, hailed the settlement as a victory over Honig and other "notorious anti-creationists". The institute considers both evolution and creation to be theories, he says, although its instructors argue in favour of creationism. ICR grants about 25 masters degrees each year in biology, geology, astrophysics and science education.

Christopher Anderson

Critics urge reform of CITES endangered list

London

NEXT week's meeting of the Convention on International Trade in Endangered Species (CITES) in Kyoto, Japan, promises to be the most acrimonious in the treaty's 19-year history.

Already on the agenda is a fight between several southern African nations and their neighbours, teamed up with the industrialized world, over proposals to modify a three-year ban on trade in elephant ivory. But informed observers say this battle is part of a larger issue that member nations must address: badly needed reform of CITES itself. The convention is performing poorly, according to leading conservation biologists, largely because it lacks appropriate criteria for

deciding which species should be included under its trade controls.

South Africa and a consortium of five southern African nations (Zimbabwe, Botswana, Malawi, Namibia and Zambia) want to remove their thriving elephant populations from Appendix 1 of CITES, a listing that prohibits international trade in any elephant products. Despite the serious problems with elephant poaching elsewhere in Africa, these countries can claim that their populations are managed effectively. They want to sell ivory from elephants that are culled annually, to raise money for wildlife conservation. Another proposal, by Richard Leakey of the Kenyan Wildlife Service, would allow countries with healthy elephant populations to

resume the less-lucrative trade in elephant skins, while retaining the ivory trade ban. Conservationists hope that this compromise can unite the divided African nations.

South Africa's request has already been endorsed by a panel of CITES-appointed experts (see *Nature* 354, 175; 1991), and expert reports on requests from the other nations will be delivered near the start of the two-week Kyoto meeting. But even if these reports are also positive, delegates from many of the 112 CITES nations are expected to block the requests. UK environment secretary Michael Heseltine, for example, has already said Britain will oppose the move.

Anticipating a tough battle, the five black southern African nations, led by Zimbabwe, are launching a full-scale offensive to rewrite the convention itself. Zimbabwe and its allies want CITES to recognize that a carefully managed trade in animal products can benefit conservation efforts. They also argue that it is now almost impossible to cancel a species' CITES listing, even if populations are no longer endangered. CITES needs 'symmetrical' criteria for including and removing species from its appendices, the southern African countries argue.

These nations also want to abolish the 'Berne criteria', by which species are given their CITES listings. Under these rules, species qualify for inclusion in Appendix 1 of CITES if they are "currently threatened with extinction". The southern African countries argue that this definition is hopelessly vague, and leading conservation biologists agree. "They're not even criteria at all," says Georgina Mace, from the Institute of Zoology in London.

Last month, at a workshop in Cambridge organized by the World Conservation Union, conservation biologists agreed that many species have been given inappropriate CITES listings. At one point, Rowan Martin, director of research at the Zimbabwean National Parks Department, challenged biologists to name a single species that has benefited from a CITES listing. There were several moments of embarrassed silence before a handful of species were put forward.

The southern African nations want to replace the Berne criteria with new definitions developed for the Union by Mace and Russell Lande, from the University of Oregon. The new definition would cover those species that have a 20 per cent chance of becoming extinct within 20 years, or ten generations (*Conservation Biology* 5, 148; 1991).

Mace believes that her criteria are not yet ready for general use; she says there have so far been no studies to test their usefulness. But she is hopeful that the

BRAZILIAN UNIVERSITIES

Soviet immigrants trigger debate

São Paulo

A TRICKLE of scientists from the former Soviet Union who are headed towards Brazil has touched off a heated debate about the wisdom of hiring them at a time when native scientists are struggling to find jobs.

The current recession has meant few opportunities for faculty of any stripe. The notable exception is the University of São Paulo, the biggest and richest university in the country which, not coincidentally, can also boast of having the most talented faculty. The university, supported by the state of São Paulo, is putting the finishing touches to contracts with 10 scientists from abroad, including eight from the former Soviet Union. One has already arrived for a two-year stay, and his colleagues — mostly engineers, physicists and mathematicians — are due this spring.

However, their hiring has been criticized by the president of the Brazilian Society for the Progress of Science, an umbrella group of scientists from various disciplines. Physicist Ennio Candotti, head of the society, believes that it is wrong for the university to hire foreigners at a time when the federal government is failing to provide necessary support for both working and would-be scientists.

Candotti pointed out that, since last May, the government has withheld funds for projects approved by an advisory body to the National Council for Scientific and Technological Development. Students who are studying abroad face similar delays. The problem came to a head earlier this month when several graduate students camped in front of foreign consulates and overseas agencies of the Bank of Brazil in cities where they are going to school to complain that they had not received their scholarship money. Their pro-

tests convinced the government to release the funds.

"The idea of bringing scientists from abroad is a very good one", says Candotti. "But we must be careful not to lose our international credibility". Candotti is also worried that foreign scientists could face the same hardships as native researchers if the Brazilian government fails to sustain its support. "Our wages are not competitive internationally," he says. Although São Paulo plans to pay the immigrant scientists some \$3,000 to \$3,500 a month, foreign scientists expected this fall at the University of Brazilia will be earning only \$1,200 to \$1,400 a month.

São Paulo's pro-rector for research, Emey Plessman de Camargo, says that his university's recruitment of foreigners goes back more than 50 years, and that "we must keep the tradition of bringing competence [to the university] from abroad". He says that the university's \$400 million annual budget can easily absorb the cost of the new faculty.

The first contingent of Russian scientists has already arrived, and is making a difference. The University of Ijuí, a small institution in the southern state of Rio Grande do Sul, has hired a small group of chemists, applied mathematicians and physicists from the Kazan Aviation Institute, a research facility that until recently has been closed to foreigners. They are currently teaching in a small town far from the country's main academic research centres.

The group are part of a continuing effort by the university to bring in foreign faculty for two-year periods. "We're very pleased with their work," says Telmo Frantz, the university's rector. "We hope that they stay."

Ricardo Bonalume

discussion will bring modern conservation biology to bear on CITES, something that has so far happened only to a "very limited" extent.

Mace also believes that CITES needs to go beyond whether a species is endangered and take into account the effect of trade on wild populations. Many endangered species are not threatened by international trade, she says, while other species that are not yet endangered are declining rapidly as a result of trade.

The southern African countries are also trying to score political points over their opponents from the developed world. One audacious proposal is for the Atlantic herring to be placed in Appendix 1 of CITES, a move that would make international outlaws of many European fishermen.

James Martin Jones, from the World Wide Fund for Nature (WWF) says the herring is now recovering from the overfishing of the 1970s, so that any such proposal should be defeated easily. But "had they gone for haddock or even cod, they could have made a strong case," he says. Nevertheless, the underlying message of the southern African countries is clear: countries that have no elephant populations of their own should not be laying down the rules for elephant conservation.

Few observers are willing to predict the outcome of the vote in Kyoto on southern Africa's elephant populations. But if all of Africa can come to a consensus, the developed world is expected to agree with its line.

Peter Aldhous

ANIMAL RESEARCH

Importer indicted

Washington

MATTHEW Block, president of Miami-based **Worldwide Primate**, a major US importer of research primates, was indicted last week by a federal grand jury for violating the Endangered Species Act. In the four-count indictment, the grand jury charged Block with attempting to ship six baby orangutans out of Bangkok, Thailand.

Orangutans are an endangered species and may not be sold or transported without a special licence. According to a statement released by the US Attorney for the Southern District of Florida, most of the smuggled orangutans, which were labelled as birds, "succumbed to the stress of their captivity and poor handling" despite an international effort to save them after they were intercepted by Thai authorities in 1990. If convicted on all counts, Block faces up to 12 years in prison and \$700,000 in fines.

Last year, Block sued Shirley McGreal, chairwoman of the International Primate Protection League, claiming that she had damaged his reputation in the course of helping the US authorities. That case is still pending. Christopher Anderson

Warning on population growth

London

THE US National Academy of Sciences and the Royal Society of London have this week published a joint statement on present global problems that gives central attention to world population growth. Among other things, the statement calls for an inter-academy conference, likely to be held in Sweden next year.

Although the frequent US reports by the National Research Council are technically opinions to which the council of the US academy assents, the Royal Society has traditionally been shy of committing itself to opinions on public policy, citing the difficulty of reaching agreement within its council. The joint statement is claimed to be an occasion without precedent.

This is also the first time the two academies have issued a joint statement on a matter of general interest, reflecting (the statement says) joint "deep concern". The statement took two years to write, and was completed only a few weeks ago. A visit last year to London by Carl Djerassi appears to have helped powerfully to crystallize the text now published.

Next year's conference will be organized by the Royal Swedish Academy of Sciences, and will probably be held in the spring. Other academies worldwide are being invited to act as co-sponsors.

The declaration links the estimated growth of the world's population of 100 million a year with the way in which human activities are producing "major changes in the global environment". If growth remains unchecked, and the patterns of human activity unchanged, according to the statement, "science and technology may not be able to prevent either irreversible degradation of the environment or continued poverty for much of the world".

Population estimates are based on the 1991 report of the UN Population Fund, which noted an acceleration of population growth since 1984. Even assuming sustained decreases of fertility towards the replacement level of 2.1 offspring per woman per lifetime, the statement says that world population may be 10,500 million in 2050, close on 90 per cent of it in developing countries.

The two academies argue that both developed and developing countries add to the pressure on the environment. Developed countries, with 85 per cent of the world's gross national product, account for most resource consumption; their emission of carbon dioxide "has the potential for altering global climate", with global consequences. Although resource consumption per head is lower in developing countries, the pressure of population and for economic growth may have similar

consequences.

Present patterns of activity coupled with population growth, the statement says, should make "even those most optimistic about future scientific progress pause". It says that some environmental changes may irreversibly damage the Earth's capacity to sustain life, noting that many species have already disappeared or are likely to do so.

The academies go on to urge that this year's UN Conference on Environment and Development should acknowledge that both human activities and population growth are essential to a consideration of a sustainable world society. They urge more "effective family planning", coupled with economic and social change, for the developing countries, and conservation, recycling, substitution and the more efficient use of energy for industrialized states. Mitigating carbon dioxide emissions is an immediate goal.

Turning to how science can mitigate these problems, the academies back (in order) "new generations of contraceptives", alternative energy sources, improved agriculture, animal and plant genetics, biotechnology and public health (including vaccines against infections such as malaria and AIDS).

Biodiversity gets a special commendation. The statement says it is important that the "nature and dimension" of the world's biodiversity should be understood. But the present rate at which species are disappearing is the greatest in 65 million years (since the Cretaceous-Tertiary extinction). The "loss of biodiversity . . . is irreversible . . . and has serious consequences for the human prospect in the future".

The statement also warns that people should not expect too much from science. It says that while research and innovation can mitigate many of the world's problems, "it is not prudent to rely on science and technology alone to solve problems created by rapid population growth, wasteful resource consumption and harmful human practices".

The statement refers to most of the issues now conspicuous on the agenda of environmentalists, some of which are controversial. But it stops short of seriously contentious questions, such as the policy of the US government on the provision of overseas aid linked with abortion and the position of the Roman Catholic Church.

More generally, the statement concludes with a commendation of "global policies . . . to promote more rapid economic development throughout the world, more environmentally benign patterns of human activity and a more rapid stabilization of world population".

John Maddox

The position of British science

SIR — The view that the international position of British science has been declining relative to other countries is supported by various science indicators, but Terence Kealey has claimed (in April 1991¹ and again in August² that British science is growing. His argument exhibits very selective use of evidence. Some of his assertions have already been laid to rest^{3–7}. However, a number still need to be addressed, particularly now that new data are available.

A recent report from the French Observatoire des Sciences et des Techniques (OST) shows that the British share of world scientific publications fell by 3 per cent between 1982 and 1988 (and by 5 per cent relative to countries of the European Communities)⁸. This is consistent with statistics from the Dutch Advisory Council for Science and Technology Policy, which reveal that, among leading scientific nations, Britain experienced the largest drop (3.5 per cent) in publication share between 1984 and 1990⁹. Furthermore, when the results are expressed in publications per capita, the relative decline is even sharper (around 15 per cent). Nor is the fall confined to one or two fields. The OST results indicate that the British publication share decreased in all fields except clinical medicine, with physical and biological sciences (down 16 per cent and 12 per cent respectively) experiencing the largest falls⁸.

There is also new evidence from citation indicators on the changing impact of British science on the international research community. Analysts at the Institute for Scientific Information (ISI, the publisher of the *Science Citation Index*) have examined trends in the average number of citations per paper earned by G7 nations over the period 1981–90. The results show that the United States, the Federal Republic of Germany, Japan and France all improved their citation impact relative to the world average, while Britain experienced the largest drop (down 3.4 per cent)¹⁰. ISI researchers have subsequently examined the UK data on journal articles in more detail and found that clinical medicine suffered the greatest fall (by 8.6 per cent), with engineering the only field to rise¹¹.

A similar picture emerges from the OST analysis of the relative impact of papers from different countries. (An impact score of 1.0 corresponds to the world average citation rate.) While UK papers continued to achieve slightly greater impact than average (scoring 1.03 in 1988), there was a steady decline

(of 4 per cent) over 1982–88⁸, a trend confirmed by ISI data¹⁰. The United Kingdom was overtaken by Germany (up 13 per cent to 1.05) during the second half of the 1980s. Fields where the UK impact score is now below 1.0 include clinical medicine (down 4 per cent to 0.98), biology (down 14 per cent to 0.92) and physical sciences (down no less than 19 per cent to 0.87). The one bright spot is again engineering (up 11 per cent to 1.16)¹¹.

Although Kealey examines Britain's record in producing frequently cited papers during 1983 and 1984 (and draws from this the general conclusion that "throughout the 1980s, the British came second"¹), he fails to look at other years. For the 300 most cited papers published in 1986 or 1987, Britain was actually only third equal with Germany (19 papers each) and some way behind second-placed Japan (23 papers). There has been a substantial decline in the UK share of such publications during the 1980s. On a whole-counting basis, the UK world-share fell from 8.0 per cent for papers published in 1981–82 to 5.3 per cent for those published in 1986–87 (the latest two-year period for which comparable figures for the three fields have been reported by ISI)¹². The decline is equally marked when fractional counting of papers involving international collaboration is employed, the UK share dropping from 7.0 per cent for 1981–82 papers to 4.7 per cent for those published in 1986–87. Most disturbingly, the UK share (again on a fractional basis) of the 100 most cited papers in physical sciences published in 1987 was only 1.3 per cent (down from 6.3 per cent in 1982), well behind the United States (71.5 per cent), Japan (11.0 per cent), France (3.5 per cent) and Germany (1.7 per cent).

How do these bibliometric statistics compare with the latest data on Nobel prizes? If one divides the past 30 years into five equal periods, in the first of these (1962–67 inclusive) the United Kingdom earned nine prizes, well ahead of France and the Federal Republic of Germany (each with four) and not far behind the United States (with 14). In both the second and third periods, British scientists won six prizes. In 1980–85, the number dropped to four, and for 1986–91 there was just one. In comparison, the 1986–91 figures for other countries include two for France, three for Switzerland, nine for Germany and 22 for the United States. If the figures for the twelve-year period 1980–91 are normalized to take into account the differing sizes of countries, Britain now lags well behind the United States, Germany, Sweden, Denmark and Switzerland in terms of Nobel prizes per head of population.

In conclusion, while it is often possible to find a few isolated indicators to support even the most unlikely of viewpoints, the overwhelming weight of evidence clearly belies Kealey's optimism about the state of British science.

BEN R. MARTIN
JOHN IRVINE

SPRU, University of Sussex,
Falmer, Brighton BN1 9RF, UK

1. Kealey, T. *Nature* **350**, 370 (1991).
2. Kealey, T. *Nature* **352**, 466 (1991).
3. Roberts, D. H. *Nature* **350**, 550 (1991).
4. Ashall, F. & Goate, A. M. *Nature* **350**, 550 (1991).
5. Collins P. *Nature* **351**, 9 (1991).
6. Humphreys, C. *Nature* **351**, 513 (1991).
7. Mulvey, J. *Nature* **351**, 513 (1991).
8. Observatoire des Sciences et des Techniques *Science et Technologie Indicateurs* 1992 (Economica, Paris, 1991).
9. Horn, T. C. M. & Langendorff, T. *Science and Technology Indicators 1991* (Advisory Council for Science and Technology Policy, The Hague, 1991).
10. *Science Watch* **2** (No. 1), 1–2 (1991).
11. *Science Watch* **2** (No. 2), 8 (1991).
12. Garfield, *Current Contents* **17**, 3–18 (1990); **18**, 3–14 (1991); **27**, 3–15 (1991).

Page charges

SIR — Y. D. Sharma (*Nature* **355**, 104; 1992) points out that some journals charge authors for processing a manuscript for reviewing, a publication cost for each page and the cost of reprints.

In our view, while a charge for the cost of reprints is reasonable, the other charges are not justified. The costs of the processing and publishing of manuscripts should be covered by the revenues generated through subscriptions and advertisements. These costs should not be charged to the authors, upon whose academic merits the success of the journal is determined. Many authors are in any case also reviewers, who are unpaid and who unselfishly donate their professional time to improve the quality of journals. Charges may reduce the potential of a journal to attract authors of papers of high academic quality. We agree with Sharma that these charges are particularly discriminatory to authors from less developed countries.

ALEXANDER K.C. LEUNG
Wm LANE M. ROBSON

Department of Paediatrics,
Alberta Children's Hospital,
University of Calgary,
Calgary, Alberta,
Canada T2T 5C7

Patent error

SIR — M. P. Bratzel's statement (*Nature* **355**, 292; 1992) that Alexander Graham Bell moved to the United States and secured US citizenship as a prerequisite for obtaining a US patent is in error. US citizenship has never been required in order to file for and obtain a US patent.

RICHARD I. MATELES

Candida Corporation,
Suite A-1706, 175 West Jackson Blvd,
Chicago, Illinois 60604, USA

Science at the atomic scale

Philip Ball and Laura Garwin

On the evidence of *Nature's* recent conference in Tokyo, technologies at the nanometre scale are now within reach. But how are they to be realized, and what form will they take?

"I CHALLENGE the notion that nanotechnology exists," said Don Eigler at *Nature's* conference* convened to discuss that topic. Coming from someone whose atom-manipulating feats serve in some eyes to define the field, this may seem ironic; indeed, each talk at the conference gave concrete evidence of our ability to observe and control matter at the atomic and molecular scale. But Eigler's point was a subtler one: he suggested that no technologies had yet come from the basic science.

This, of course, begs the question of when science becomes technology. But what was made abundantly clear by speakers ranging from solid-state physicists and electronics engineers to organic chemists and materials scientists, is that science at the nanometre scale is full of potential for practical exploitation.

For a start, there is already evidence (or what John Hopfield (Caltech) calls an "existence theorem") for a working molecular technology, in the form of biology. Mechanical, electronic and computational operations can, it seems, be conducted efficiently by relying on molecular interactions; moreover, the machinery for such processes can to some extent assemble itself from the basic components. What remains for chemists is to identify which are the guiding principles that one might usefully mimic in making artificial molecular assemblies. In materials science, too, there is much to be learned from the way that biology contrives to engineer superior properties by control of microstructure and growth.

For more conventional electronics technology, based on semiconductors such as silicon or gallium arsenide, the challenge of the nanometre scale is not so much that of constructing small enough devices, as of coping with the new physics that becomes manifest on entering the domain of quantum mechanics. But this new physics has also been turned to advantage, not only in providing the basis for novel devices but as a fascinating field for basic research.

Novel devices

Not surprisingly, much of the impetus towards nanotechnology has come from the electronics industry. In the past 40

years, the characteristic size of a transistor has decreased from 1 cm to less than 1 μm ; there are now commercial integrated-circuit chips containing more than a million devices. At this scale, macroscopic principles of electronic circuit design are still valid, but a further 10–100-fold reduction, which could be achieved in one or two decades, will bring individual devices to a size at which the quantum-mechanical, wave-like nature of the electron can no longer be ignored.

This 'mesoscopic' realm, in which an electron's behaviour is dictated as much by the structure through which it travels as by bulk material properties, has been explored by means of engineered structures that confine electrons to progressively fewer dimensions. As Hiroyuki Sakaki (Univ. Tokyo) pointed out, quantum confinement in two dimensions has existed for 20 years in the form of ultra-thin semiconductor films deposited by molecular-beam epitaxy or chemical vapour deposition. As the thickness of such a film is reduced, the electron energy for motion normal to the layer is quantized into discrete sub-bands — just as only certain frequencies of sound are allowed to resonate in a wind instrument. For a layer as thin as ~ 10 nm, the energy spacing between bands is large compared to the characteristic thermal energy $k_B T$, even at room temperature, so that almost all electrons are in the ground-state sub-band. This quantization already forms the basis for practical devices, such as the resonant tunnelling diode and quantum-well laser.

Beyond the quantum well lies the one-dimensional quantum wire. The further reduction in dimensionality brings with it markedly higher electron mobilities, because elastic scattering of electrons by impurities is limited to nearly exact backward scattering, which has very small probability. Inelastic scattering by optical phonons, the dominant process at high temperatures, can also be suppressed if the ground-state bandwidth is narrower, and the first energy gap wider, than the phonon energy, so that absorption or emission of phonons is prevented. In practice, this means that quantum coherence effects can be observed at room temperature, provided that the wire width is less than ~ 20 nm. This has allowed the fabrication of 'quantum interference devices', in which

external signals control the interference of coherent electron waves, to increase or reduce electron transmission. Quantum wires can also support high electron velocities, making an array of quantum wires an attractive option for more conventional high-speed devices.

New physics

The same high electron mobility that makes mesoscopic conductors attractive for device applications has made them a veritable playground for those interested in the basic physics of electron transport. When the dimensions of the conductor are smaller than the electron mean free path, the electrons travel 'ballistically', undisturbed by interactions with phonons or impurities. When these dimensions are also comparable to the electron wavelength (strictly speaking, the Fermi wavelength), quantization effects are observed.

Bart van Wees (Univ. Groningen) described the completely unexpected discovery of conductance quantization in a quantum point contact (B. J. van Wees *et al. Phys. Rev. Lett.* **60**, 848–850; 1988). The quantum point contact creates a constriction of variable width through which electrons confined to a two-dimensional semiconductor layer can pass; the unexpected finding was that the conductance of this constriction exhibits plateaux as a function of its width. This behaviour is now understood to arise from the quantization of transverse momentum in the constriction: each increase in width by half the electron wavelength allows an additional one-dimensional sub-band, or 'channel', to be added to the conductance, increasing it by $2e^2/h$. (Those familiar with the quantum Hall effect will recognize $2e^2/h$ as the spacing of plateaux in the Hall conductance (ignoring spin splitting); van Wees explained that the two effects are indeed related, as the variation of conductance with magnetic field can also be explained in terms of the occupation of quantized channels.) The surprise was that this behaviour, which had been predicted to occur in an ideal quantum wire in which scattering is absent, can be observed in a real experimental system.

A different type of quantization — that of electric charge — also rears its head as circuits shrink. Michel Devoret (Centre d'Etudes Nucléaires de Saclay) described how, at low temperatures and

* Nanotechnology: Science at the Atomic Scale, Tokyo, 27–28 January 1992.

in sub-micrometre-scale metallic circuits, it is possible not only to detect single electrons but to manipulate them. The phenomenon of single-electron tunnelling (SET) arises from the interplay of electron delocalization and the Coulomb interaction: an electron is allowed to tunnel across a narrow insulating barrier to a region of low capacitance, C , only if it can supply the requisite electrostatic charging energy $e^2/2C$. At low temperatures and/or very low capacitances, this energy can be made large compared with thermal fluctuations, so that passage through the barrier (lifting of the 'Coulomb blockade') can be controlled reliably by an external gate voltage. This is the basis for the 'electron turnstile' (L. J. Geerligs *et al.* *Phys. Rev. Lett.* **64**, 1861–1864; 1989), in which a radio-frequency gate voltage 'clocks' electrons one by one around a circuit, with an accuracy of 10^{-3} . With sub-micrometre circuit elements, the characteristic capacitance is about 10^{-15} farad and single-electron tunnelling can be seen at temperatures below about 0.3 K; if the dimensions could be reduced to the sub-nanometre scale, with capacitances of about 10^{-18} farad, then SET could be observed at room temperature.

Just as two Josephson junctions can be arranged in parallel to make an exquisitely sensitive device for measuring magnetic flux (the superconducting quantum interference device, or SQUID), two normal-metal tunnel junctions can be combined in an exactly analogous circuit to yield a charge-measuring device termed the 'SET transistor'. With this device, one can observe single tunnelling events, and monitor the charge on an electrode with sub-single-electron accuracy. In principle, SET (in the shape of the electron turnstile) provides an exact relationship between frequency and current, just as the Josephson effect ties together frequency and voltage, and the quantum Hall effect links voltage to current. Devoret hopes that SET may someday "close the metrological triangle" by providing a very accurate current standard, but notes that this would require a turnstile accuracy of 10^{-8} , instead of the present value of 10^{-3} .

At first sight, the SET transistor might seem to be the fulfilment of a nanoelectronic dream, with only one electron needed for each bit of information. But Devoret points out that as the transistor has no voltage gain, it is unlikely to be of interest to engineers. The same can certainly not be said of its analogue, the SQUID. In a presentation that showed just how quickly new science can be turned into commercially viable technology, John Clarke (Univ. California, Berkeley) demonstrated that high-transition-temperature oxide superconductors, the

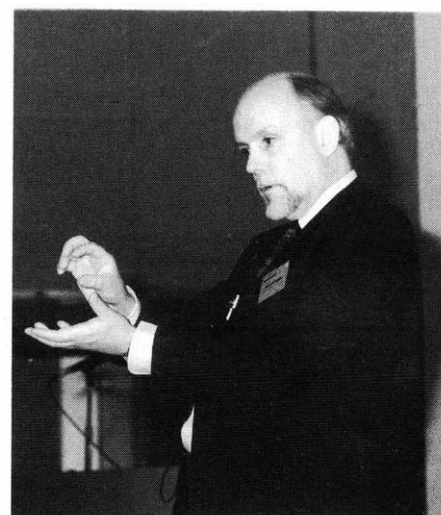
first of which was discovered only in 1986, already provide the essential components for an electronics technology. As with semiconductor electronics, the fabrication of superconducting devices relies on the controlled epitaxial growth of thin films, followed by photolithographic patterning. Individual Josephson junctions have been patterned with widths as small as 0.2 μm , so that device sizes of a few hundred nanometres seem within reach. Closer to the market, Clarke described an 'integrated magnetometer' comprising a SQUID and a flux transformer (for increased sensitivity), made by depositing eight epitaxial layers on the same substrate; as yet it works only at 75 K and below, but there is no reason why this should not be increased to ~ 90 K.

Another lesson that would-be nanotechnologists might learn from the high- T_c superconductor experience is that any emerging technology will have to compete with whatever technology exists already — in the case of digital electronics, this means silicon chips. It looks likely to be a long time before a Josephson computer will be able to compete with silicon, but high- T_c superconductors have nevertheless already found several niches outside computing, notably in far-infrared detectors, microwave components and magnetometers

Making nanostructures

The wealth of new devices and new physics discussed above suggests that we are not yet limited, either in technology or basic science, by our ability to fabricate small structures. Defining 'nanostructures' as structures with dimensions less than 0.1 μm , Alec Broers (Univ. Cambridge) inferred from past trends that the mainstream electronics industries would not be needing nanostructures for more than 10 years, yet present-day electron-beam lithography can already be used to fabricate useful devices measuring ~ 20 nm. (The resolution of this apparent paradox lies in the fact that electronics technology comprises not just individual devices, but literally integrated circuits; it is the need of each successive generation of chips for new circuits, new materials and new design concepts that seems to limit the speed of progress.) But even if electronics companies are not forcing the pace, there is surely a demand from research physicists for ever-smaller structures. So what are the limits on nanolithography?

Electron optics can produce beams of about 0.2 nm in diameter, but the smallest practical structure made so far is 100 times larger than this. As Broers explained, this limit arises from the use of polymers as 'resists' in the electron-beam lithographic process: where the polymer has been exposed to the beam, it can be



Richard Smalley encapsulating fullerene research.

preferentially etched away, leaving a pattern for subsequent deposition of circuit elements on the substrate. But although it is possible to write a line as thin as ~ 1 nm on the resist, it will not be etched successfully; the thinnest lines that can be developed properly are ~ 10 nm wide. This behaviour is not completely understood, but it seems that inelastic interactions between electrons in the beam and resist molecules — for example, secondary electron excitation and subsequent straggling of the secondaries — give rise to a smearing-out of the electron dose with a characteristic width of 12 nm. This behaviour seems to characterize all spin-on polymer resists, irrespective of molecular size.

Two techniques have been used to overcome the 10-nm resist limit, but each has its problems. In direct structuring sublimation, the electron beam is used not to expose a resist, but to sublime a metal halide or oxide substrate. This technique has been used to drill holes 2 nm in diameter and 4 nm apart, but has not yet created any useful structures. In direct-exposure lithography, the beam is used to write a pattern on SiO_2 , with the effect of enhancing the etch rate in hydrofluoric acid. This technique has been used to create slots 10 nm wide and 15 nm apart, which can then be replicated in silicon by reactive ion etching, using the patterned SiO_2 as a mask. The difficulty here, however, is the rather small difference in etch rate (about a factor of three) between exposed and unexposed regions of the SiO_2 , giving rise to slots with sloping sidewalls.

The higher resolution achieved with these two techniques lends support to the idea that what is needed is an image-forming process that is activated only by the high-energy primary electrons, rather than by the low-energy (20 eV) secondaries. Breaking a bond in a

polymer such as polymethylmethacrylate requires only 5 eV; defect formation in SiO₂ requires about 30 eV. "If we had something that requires 100 eV", said Broers, "we would have better resolution." This plea prompted Richard Smalley (Rice Univ.) to make the first of several suggestions of novel uses for the buckminsterfullerene molecule: why not use the electron beam to crosslink molecules in a thin film of C₆₀? This might well take a lot of energy, and the unpolymerized C₆₀ could be dissolved away with benzene.

Carving versus constructing

The struggle to extend lithographic techniques to smaller scales raises the question of how far it is worth pursuing a 'top-down' approach to nanofabrication. As Hopfield put it, "Using molecular-beam epitaxy or lithography for nanoelectronics seems increasingly like making a suspension bridge by carving it out of a large block of steel." The opposite approach is to build nanostructures atom by atom, a capacity that is now available courtesy of the scanning tunnelling microscope (STM).

The sheer joy of being able to move single atoms pervaded Don Eigler's (IBM Almaden Research Center) presentation. He told how a problem — is the STM tip perturbing the system? — was converted by some handy lateral thinking into a new capability. It turned out that the force exerted by the STM tip on a xenon atom adsorbed on a metal surface could have a surface-parallel component large enough to pull the xenon atom across the surface, while having a surface-normal component small enough to leave the atom bound to the surface. Eigler's first feat was to construct a linear chain of seven xenon atoms on a nickel (110) surface; his colleague C. P. Lutz then did the same with platinum atoms on platinum — a trickier task than for weakly bound xenon. Eigler looks forward to the day when we will be able to use the atomic force microscope to do similar things with metal atoms on an insulating surface — building circuits atom by atom to study electron transport in exceedingly small structures.

Not content with a flatlander's approach to atomic manipulation, Eigler and others have since learned to pick up atoms with the STM: this allows atoms to be carried over step edges, and provides in principle for the fabrication of three-dimensional structures. Eigler's two-terminal "atom switch", in which the reversible transfer of a xenon atom from the STM tip to a nickel surface modulates the conductance between tip and surface, can be thought of as the prototype for an atomic-scale electronic device. But Eigler is the first to point out

that this 'atomic-scale' device has a roomful of equipment attached to it. In fact, it is not clear whether the atom switch would work with nanoscopic leads, which might not provide a sufficient heat sink to prevent the device from destroying itself.

Eigler stressed that the mechanism that causes the xenon atom to move back and forth from tip to surface is not yet understood; such general questions of mechanism are being addressed systematically by Masakazu Aono (RIKEN) in the ERATO "Atomcraft Project". Perhaps as an incentive to solve the difficult problems standing in the way of reliable fabrication, Aono poses the question, "If you have a monolayer of silver atoms on silicon, and remove one out of three silver atoms, will you get a SET transistor working at room temperature?" Even if not, it is clear that the rewards will be enormous. In fact, Aono's colleague H. Nejoh (*Nature* **353**, 640–642; 1991) has claimed to have seen single-electron tunnelling from an STM tip to a liquid-crystal molecule at room temperature. This prompted Devoret to suggest that the ubiquitous C₆₀ might be an even better molecule to use, as it would be a nearly ideal spherical capacitor.

Aono described a systematic set of experiments from which he concluded that field evaporation of ions from both tip and sample is the cause of the atomic-scale ditches and bumps created by scanning an STM tip over the reconstructed surface of silicon known as 7×7. The hope is that such experiments will ultimately provide the bedrock of fundamental understanding that will have to underlie a successful manipulation technology.

Another essential requirement for any type of nanofabrication system is the ability to position a tool (whether it be an STM tip, an electron beam, or even a molecule) with nanometre accuracy. Shoichiro Yoshida (Nikon Corp.) demonstrated that, even if there is as yet no technology based on nanostructures, we already have a technology with nanometre precision. In particular, Yoshida's ERATO "Nano-mechanism Project" has developed a fine position control for 1-nm positioning, using a rolling ball guide, and a laser interferometer for measuring nanometre displacements, which by using two different wavelengths reduces errors of 30–40 nm caused by atmospheric turbulence to a few nanometres.

From self-assembly to devices

Building molecular devices and assemblies by manipulation of individual molecules is, with the STM's new-found versatility, not beyond our means, but it is likely to be prohibitively slow. At the

nanometre scale, it will be much more efficient to let chemistry do the job — that is, to design molecules that will assemble spontaneously into useful structures. Mutsuyoshi Matsumoto (National Chemical Lab., Tsukuba) and George Gokel (Univ. Miami) provided elegant examples of the extent to which self-assembly can now be controlled. In neither case were the basic structures hot off the drawing board: the ordered, lamellar arrays of amphiphiles described by Matsumoto were studied by Irving Langmuir and Katharine Blodgett in the 1910s, while Gokel's bilayer vesicles have a history as old as the cell. But clever chemistry has allowed functionalities to be built into these structures of the sort that could generate many new applications.

Matsumoto described Langmuir-Blodgett (LB) films whose electrical conductivity, potentially substantial by virtue of the TCNQ-based charge-transfer complexes incorporated in the amphiphile head groups, is photochemically switchable (H. Tachibana *et al.* *J. Am. chem. Soc.* **111**, 3080–3081; 1989). This immediately invites speculative applications for energy conversion and signal transduction. The amphiphiles are designed on a modular basis: as well as the charge-transfer 'working' unit, they contain a switching unit in the 'tail', linked to the head-group working unit by a linear alkyl 'transmission' unit. The switch is an azobenzene group, in which the –N=N– double bond may adopt either a *cis* or a *trans* conformation. The *trans* form is favoured at room temperature, yielding a more or less linear molecule which will form an ordered two-dimensional array when a monolayer on the surface of water is compressed in an LB trough. Multilayer, lamellar stacks can be deposited from this apparatus onto a solid substrate by a simple dipping procedure. As the head groups lie side by side in a periodic fashion, these deposited LB films show an appreciable in-plane conductivity. But irradiation with ultraviolet light switches the conformation to *cis*, leaving the molecules kinked and altering the arrangement of head groups; the conductivity drops as a result. Matsumoto showed that a similar photoinduced switch in a monolayer film can cause reorientation of molecules in a smectic liquid-crystal film above it; this might form the basis of a new method for switching liquid-crystal displays.

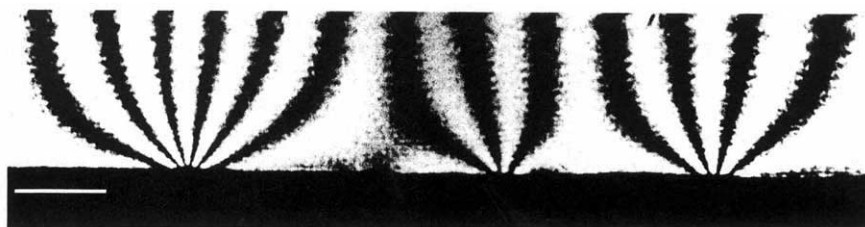
Switching behaviour also lies at the heart of Gokel's self-assembling structures. A reduced cholesterol (nature's own membrane stiffener) attached to ferrocene (an Fe²⁺ ion sandwiched between two cyclopentadienyl rings) constitutes a somewhat unusual kind of amphiphile, but these molecules will

Seeing the unseeable

The ability to move atoms and molecules with the STM has perhaps diverted attention from the equally wonderful possibilities for 'seeing', and more broadly, sensing, offered by this microscope and its relations. Heinrich Rohrer (IBM Zürich Research Laboratory) surveyed the broad sweep of these possibilities, pointing out that the range of possible interactions between tip and sample (van der Waals, adhesive, magnetic and electrostatic, to name a few) offers both opportunities and potential problems. In principle, one can isolate any property of interest and measure it locally: under the broad heading of magnetism, for example, one can use the magnetic valve effect (in which the tunnelling current is sensitive to the electron polarization), magnetic force microscopy (with a magnetized tip), or magnetic resonance microscopy. Any of these techniques can also be combined with spectroscopy — for example, one might inject polarized electrons from the tip, and measure the emission of polarized photons. But Rohrer cautioned that the sheer number of interactions makes it hard to be sure that one is really looking at the interaction of interest, and one must also ensure that the probe itself doesn't interfere with the experiment one is trying to perform.

Akira Tonomura (Hitachi Advanced Research Laboratory) reminded the

audience that the STM is not the only way to create atomic-scale images. The availability of coherent electron beams, produced by field emission from a pointed tungsten tip, has led to the technique of electron holography, for both high-resolution imaging and electron interferometry. Tonomura illustrated the use of reflection interferometry to measure surface topography with a resolution of less than 0.01 nm. Like the STM, however, the technique can do far more than simply sense topography. Thanks to the Aharonov–Bohm effect, the phase of an electron wave is sensitive to the presence of an electromagnetic field: the fringes in an interference micrograph appear along magnetic lines of force, with a separation of one flux quantum between adjacent fringes. Following the use of electron holography to demonstrate definitively the existence of the Aharonov–Bohm effect, Tonomura has used the technique to image flux lines in superconductors — providing direct evidence for flux-line pinning, and other aspects of flux-line dynamics induced by thermal excitations or interactions with the supercurrent. Having seen Tonomura's videotape of flux lines dancing across the surface of a superconductor (see figure), one is inclined to believe him when he says, "with this technology, even thought experiments will be possible". □



Electron interference micrograph showing bundles of flux-lines emerging from the surface of a superconducting lead film. Scale bar, 2 μm .

nevertheless form bilayer vesicles in just the same way as phospholipids. This propensity is, however, determined by the redox state of the iron — electrochemical oxidation of the Fe^{2+} to Fe^{3+} causes vesicles to form, while reduction breaks them down again (J. C. Medina *et al.* *J. Am. chem. Soc.* **113**, 365–366; 1991). Gokel has also demonstrated redox-switchable binding of metal ions in a ferrocene derivative in which the two cyclopentadienyl rings are bridged by several tertiary diamine loops (nitrogens joined by aliphatic chains), leaving a rigid cavity between the nitrogen donor atoms. The bound metal ion sits next to ferrocene's Fe^{2+} , and is therefore expelled when the iron is oxidized to give a positively charged

ferrocenium species.

One might also regard as a kind of mechanical switching the structural changes that binding of substrate to receptor can initiate. Particularly suggestive of this mechanical response was Gokel's 'cleft-like' receptor in which the rotational freedom of ferrocene's aromatic five-carbon rings around the iron 'ball bearing' is exploited to build receptors with arms that swing around to lock their substrates in place. Of an entirely different nature is the mechanical switching of toroidal molecules threaded on a linear chain — the so-called rotaxanes. Fraser Stoddart's group in Sheffield has proposed that their rotaxanes, comprising cyclic bipyridinium units threaded on linear hydroquinols or vice versa, might

be assembled into a 'molecular abacus'; a variant presented by Gokel was a cyclodextrin on a hydrocarbon chain capped with ferrocene and a sulphonate group.

Masuo Aizawa (Tokyo Inst. of Technology) suggested that mechanical switching also lay behind the transfer of conformational change between two proteins covalently linked by a kind of molecular transmission shaft. Binding of a calcium ion to calmodulin converted from an inert to an active form the phosphodiesterase enzyme to which it had been attached by a covalent bond.

To some extent one might regard these molecular-based switches as components in search of a device. But as an example of a practical device that relies on engineering and interactions at the molecular scale, the biosensor provides arguably the best example currently on the scene. The earliest glucose sensors, developed in the 1960s, were little more than electrochemical oxygen sensors that depended on the immobilization of the enzyme glucose oxidase at the electrode surface. But Aizawa discussed possibilities for improving signal transmission at the molecular level. He showed that the electrical connection between the enzyme and the electrode can be enhanced by attaching to the electrode surface molecular wires (in this case conjugated polypyrrole molecules), the free ends of which may find access to the redox site on the enzyme.

Materials design

Fullerenes, inevitably, surfaced in many contexts: C_{60} is perhaps the perfect nanoscale construction material, a structural unit itself only marginally smaller than a nanometre. But Richard Smalley considers the graphitic microtubules reported recently (S. Iijima, *Nature* **354**, 56–58; 1991) to be of potentially equal importance; calculations indicate that some may be semiconducting and that their perfect crystallinity will make them the strongest fibres known. It is far from nanotechnological oversell to imagine welding these girders into rigid superstructures with an electron beam (perhaps having first arranged them with the STM?). The hierarchical structure of these frameworks, the basic elements of which are themselves built from the graphite 'chicken-wire' mesh, should imbue them with some of the advantages that hierarchy bestows on natural substances such as bone, an issue addressed by Paul Calvert (Univ. Arizona).

Among these superior properties, said Calvert, are toughness, damage tolerance (local failure need not lead to global breakdown), repairability and ease of further growth. Bone provides perhaps the classic example: the porous

ity, extending over several length scales, not only allows transport of fluids but also gives flexibility and high strength-to-mass ratios. Growth of bone and similar biomineral materials such as tooth enamel and shell is generally a template-directed process that allows for structural control at the micrometre level (if not finer still). Biomineralization takes place typically either in a polymer matrix (as is the case with bone) or through the guidance of a membrane, and can lead to the rapid and efficient production of extraordinarily complex shapes such as the delicate platelets of calcium carbonate found in coccoliths or the fullerene-like silica cages of diatoms.

To mimic these growth processes requires an *in situ* precipitation technique in which particle size, shape and crystallographic orientation can be controlled. In biological systems, acidic proteins act as crystal growth modifiers that can, according to the conditions, either inhibit or promote growth. Calvert suggested that treating surfaces using, for example, photolithographic methods might provide a route to localized and patterned precipitation, while drawn or brushed polymer films have already been shown to be efficient at controlling the orientation of crystallization (J. C. Wittman & P. Smith, *Nature* **352**, 414–417; 1991). But to achieve anything remotely like true mimicry of biomineralization, Calvert pointed to the need for new synthetic polymers that can rival proteins in their ability to provide a guiding matrix.

That nature's composite materials can achieve what monolithic ones cannot is no recent revelation. The essential attraction of composites lies with their potential to combine the strength of homogeneous, dense monolithic materials with the toughness of coarse-grained, inhomogeneous solids, in which weak links can absorb energy without catastrophic failure. Past attempts at making composites that combine these attributes have centred around the use of fibres and whiskers of micrometre dimensions embedded in a matrix. But Koichi Niihara (Osaka Univ.) showed how the fabrication of composites from elements of nanometre size is leading to far superior properties.

One variant of this approach incorporates nanometre-sized clusters either within or between larger grains of another phase. Intra-grain nanocomposites in which silicon carbide or nitride nanograins are dispersed in oxide matrices such as MgO and Al₂O₃, for example, can be prepared from powdered starting materials using standard metallurgical techniques — ball milling, hot-pressing and sintering. The improvements in properties relative to corresponding microcomposites are striking:

both toughness and strength can be increased by up to a factor of four, while the maximum operating temperatures are increased by 400 to 800 °C. These micro/nano composites provide another illustration of how a hierarchical structure can make for improved materials.

An alternative type of nanocomposite, consisting entirely of nanoscale grains, may exhibit entirely new mechanical properties compared with monolithic materials. A nano/nano-composite of this kind formed from silicon carbide (SiC) and silicon nitride (Si₃N₄) grains less than 500 nm in size will become ductile, or 'superplastic', at temperatures above ~1,600 °C, and can therefore be



Don Eigler, with a row of xenon atoms in the background.

readily moulded into components that retain both hardness and toughness at lower temperatures. These non-oxide composites are formed by hot-pressing of powders in which the grains have been put together atom by atom via chemical vapour deposition. It appears that the morphology and size of the silicon nitride grains can be controlled by the volume fraction of nanoscale silicon carbide particles in a manner that recalls inhibition or promotion of crystallite growth in biomineralization: at low SiC fractions, the SiC particles act as nucleation sites for the formation of large, elongated Si₃N₄ grains, leading to high toughness, while larger SiC fractions promote the growth of the equiaxed, nanoscale Si₃N₄ grains necessary for superplasticity.

Optical effects

Rather than considering the behaviour of nanoscale clusters *en masse*, Louis Brus (AT&T Bell Labs) described how their finite size affects the properties — electronic and optical — of individual clusters. When they contain only a few hundreds or thousands of atoms, one might expect the electronic structure not yet to have acquired the truly band-like character of the bulk solid, so that a description in terms of molecular orbitals

may remain valid. In semiconductors, it is the separation between the highest occupied molecular orbital (HOMO) and the lowest unoccupied orbital (LUMO) (which in a dimer translates as the splitting between bonding and antibonding orbitals) that ultimately metamorphoses into the bulk band gap. For clusters in the intermediate size regime, one might expect still to find discrete HOMO and LUMO levels, with an energy gap greater than that between the fully broadened bulk bands and therefore a blueshift of the absorption spectrum.

To study these effects, a highly monodisperse range of cluster sizes is needed.

The various ways in which this might be achieved include the precipitation of clusters in zeolites, porous glasses, vesicles and gels; but Brus showed that the use of inverse micelles as self-assembling 'cluster moulds' is particularly effective and convenient. In non-aqueous phases, these spherical assemblies of amphiphiles can encapsulate in their hydrophilic interiors small pools of water within which clusters of the semiconductor cadmium selenide can be precipitated with diameters of typically 30–40 Å. These particles can be stabilized against flocculation by giving them an organic coat of phenyl groups attached to selenium.

The hope is that these clusters may constitute 'quantum dots' — zero-dimensional systems for which the electronic levels are quantized. Spectroscopic studies do show the expected blueshift of absorption peaks relative to the bulk material, but the peaks are broadened, in part by the size distribution. Hole-burning spectroscopy can be used to eliminate this 'inhomogeneous' broadening, allowing the spectra of clusters of a single size to be measured.

But even these individual spectra fail to show ideal 'quantum dot' behaviour. The reason for this is that an appreciable proportion of the atoms in the particles are at or near the surface, so surface electronic states exert a strong influence. In particular, they can act to localize and trap charge carriers, leading to bleaching of the absorption. Brus suggested that chemical tailoring of the cluster surfaces might provide a way to suppress this influence of surface states.

Conjugated polymers in the bulk state provide semiconducting materials whose electronic and optical properties can be exquisitely tuned by chemical modification. Variation of the band gap of poly(*p*-phenylene vinylene) (PPV) by attaching appropriate side groups to the polymer chain, for example, can provide a means for moving the absorption maximum across the optical spectrum. Richard Friend (Univ. Cam-

bridge) described how this chemical handle on the electronic properties can be exploited to create polymer-based light-emitting diodes with a range of colours. Injection of electrons into the conduction band, and holes into the valence band, of a PPV film leads to the formation of mobile radical ions (polarons) on the polymer backbone. Oppositely charged polarons can combine to form excited states (polaron excitons) in the energy gap, providing the potential for new optical transitions. Radiative recombination of these excitons leads to luminescence from the polymer films.

Friend has now extended the efficiency and processibility of these electroluminescent systems through a *tour de force* of nanoscale band-gap engineering which makes use of block copolymers. These incorporate stretches of different monomer chains (generally methoxy derivatives of *p*-phenylene units) along the polymer backbone, which can be made either conjugated (semiconducting) or non-conjugated (insulating) by chemical modification. Trapping the excitons in one-dimensional quantum wells bordered at each end by insulating sections of the chain prevents their migration to quenching sites such as defects, which otherwise causes a reduction in the luminescence efficiency; enhancements by factors of up to 30 are possible (see next week's issue of *Nature*). Moreover, patterning of these copolymer films is possible by lithographic means: masking regions of a conjugated/non-conjugated (orange) film with aluminium and treating exposed areas with acid to eliminate methoxy groups will, for example, write yellow (fully conjugated) patterns into the unmasked portions.

Molecular computing

Molecular biologists might shrink from suggestions that they too are engaged in a form of nanotechnology, but there is no question that the process of translation from the four-symbol nucleic acid code to the twenty-symbol protein alphabet is a computational operation at the molecular level (John Hopfield). Often praised for its fidelity, this process actually tolerates an appreciable error rate — a mistake every 10,000 or so operations — by virtue of massive redundancy: so many protein molecules are synthesized that a wrong one here and there (which will generally be non-functional) does not matter.

This example of a biological symbol-conversion process serves to provide Hopfield's "existence theorem" for molecular-scale computing — a demonstration that in principle it can be done. But most computational problems require more complex logical operations, such as those found in neurobiological processing. The behaviour of neural net-

works, both real and artificial, is of a more collective and ultimately macroscopic character than anything in molecular biology. A collective response, made possible by a high degree of parallelism, lowers the accuracy requirements for individual processing steps.

Can one contemplate an artificial, molecular-scale computer that could compete with the complexity of the brain (roughly 10^{11} neurons linked by 10^{14} synapses)? Aizawa demonstrated that the growth of nerve cells on an indium tin oxide substrate can be directed by applied electric fields, raising the intriguing prospect of synthetic, biologically based networks; but it is hard to see how a third dimension could be added or interfacing to VLSI technology achieved. More immediate was Hopfield's suggestion for constructing parallel networks using wires fabricated by conventional VLSI techniques to connect 'synapses' made from novel (and in the fashionable jargon, intelligent) materials such as copper salts of the organic acceptor compound TCNQ, chalcogenide glasses or tungsten bronzes. The conductivity of these materials is field dependent and hysteretic, providing the possibility of read, write and erase operations.

Although they would rely on molecular-scale engineering, such systems are nevertheless essentially macroscopic. In the way of truly molecular devices Hopfield envisages substantial hurdles. The 'molecular shift register' (J. J. Hopfield *et al. Science* **241**, 817–820; 1989), for example, exhibits behaviour that might reasonably be characterized as device-like. In this four-component molecular assembly, absorption of light by the first component, a chromophore, sends an electron tumbling down an energy-level cascade on the successive components. One could imagine the simultaneous shift of a whole bank of electrons if the four-part units could be linked into a polymeric chain. But Hopfield cautioned against considering this to be a real device. How one might process and interface such a polymer is not clear, and it lacks the essential ingredient of an error-correction mechanism. The latter requires some form of collective behaviour. The occurrence of cooperative two- and four-electron oxidations in biology, however (for example, in the action of haemoglobin), shows that ultimately this may not be too much to ask.

Does nanotechnology exist?

Hopfield characterized the field of molecular nanocomputing as suffering from "an excess of imagination and a deficiency of accomplishment"; can the same harsh assessment be made of nanotechnology more generally? Paul Horn (IBM Thomas J. Watson Research

Center), whose duties for IBM include "research, technology definition and early development of advanced logic and memory products", echoed Eigler in saying that, while the technical capabilities described at the conference provide "wonderful tools for science", he did not expect any of these to make a real impact on mainstream electronics technology ("the electronics you have in your home") in the next 25 years. As Eigler emphasized, and those working in high- T_c superconductors have learned, it is important to distinguish between what is physically possible, what is technologically realizable and (in the face of competition from well established technologies) what is economically advantageous. For example, Eigler admitted that although he could store the equivalent of a year's worth of IBM's information storage products on a one-inch wafer, it would take him the age of the Universe to write the information.

Part of the problem is that a working device, of any kind, is only the first tentative step towards a 'technology'. Supposing one has a molecular solar-cell, for example, how does one then go about collecting the electricity that it generates? If a molecule signals its recognition of another molecule by changing its shape, how can this signal be processed? Even in the more conventional world of semiconductor electronics, how does one get a useful amount of current out of a few-electron quantum device? The problem of providing interfaces between the nanoscopic and macroscopic worlds is clearly not trivial.

But as Hopfield reminded the audience, electronics isn't everything: a molecular device that behaved like a chloroplast (for carbon fixation) might be very useful, and would not require any 'leads' to connect it to the outside world. Far from an excess of imagination, our problem may be that we have not yet learned to think in the new ways that will be required to make the best use of these novel (proto)-technologies. In this sense, the lessons of nanotechnology extend beyond the very small, to encompass the general process of learning from biology and chemistry, and then going beyond what is observed in the natural world, to deliberate engineering.

It has been said of technological progress that we always overestimate what we can do in two years and underestimate what we can do in twenty. Horn's high-school science teacher taught the class that they would never see an atom. In the face of these lessons, it would seem foolish to deny that our dreams may come true sooner than we think. □

Laura Garwin and Philip Ball are, respectively, Physical Sciences Editor and Assistant Editor of *Nature*.

Melodrama in research publication

The excitement of a rapidly advancing research field may be swamped by the confusion generated by the competition of researchers (and journals) to be the first in print.

MYOTONIC dystrophy is the most common form of muscular dystrophy, affecting one in 8,000 among the people in whom the incidence has been carefully studied. It is genetically determined, indeed dominantly so. But in contrast with many other inherited diseases, the symptoms of myotonic dystrophy may make their first appearance at any age between infancy and middle life.

There are also great variations in the severity of the disease, provoking questions such as why throwing a kind of genetic switch should have an uncertain outcome. More puzzling still is the observation that, in families carrying the aberrant gene, the symptoms of myotonic dystrophy should become more severe as generations succeed each other. Clinical geneticists have a word for that: "anticipation".

These curious features of myotonic dystrophy were unexplained until a few weeks ago, when this journal published a group of three letters describing attempts to identify and characterize the aberrant gene, known for some time to be located on human chromosome 19 (Harley *et al.*, Buxton *et al.* and Aslanidis *et al.* 355 545, 547 and 548; 6 February 1992). All three reports came to the same conclusion: the myotonic dystrophy gene is closely linked with a region of DNA that is unstable in the sense that its length can differ between a person and his or her offspring.

At the same time, it became plain that the length of this variable region of DNA is positively correlated, in those with myotonic dystrophy, with the severity of the symptoms. So the question naturally arises of whether the aberrant gene differs from the normal simply in the length of this variable region. That, of course, complicates textbook notions of what aberrant genes are like: mutations consisting of the replacement of one nucleotide by another are the simplest, deletions of a group of three nucleotides (leading to the omission of a single amino acid from the product protein) the next simplest, and so on.

Aberrancy that consists of variability in the length of a piece of DNA is clearly in a different category, provoking several interesting questions. What does the variability of length consist of, and how does it arise? By what mechanism does the supposed product protein cause the symptoms of myotonic dystrophy? And so on.

For what it is worth (which may be a great deal), similar questions have previously arisen. Thus the origin of the X-linked genetic disease known as Kennedy's

disease, a form of motoneuron disease, was shown last year (La Spada *et al.*, *Nature* 352, 77; 1991) to consist of a doubling of the length of a stretch of the X-chromosome which normally consists of an average of 21 repeated nucleotide triplets CAG (for cytosine, adenine and guanine). The consequence is an extra stretch of repetitive glutamic acid in the structure of the cell-membrane receptor for androgen.

The appearance of extra nucleotide triplets ("expansion" is the new word) in the X-chromosome is also linked with the condition known as fragile X-chromosome (one of the commonest forms of genetic mental defect), but in that case the extra triplets consist of CGG (Oberlé *et al.* *Science* 252, 1,097; 1991).

So there are good reasons why people should be excited by the discovery that myotonic dystrophy is somehow linked with the phenomenon of genetic expansion. And it is far from irrelevant that myotonic dystrophy is the most common cause of muscular dystrophy. But does this sense of excitement justify the melodrama lavished on the problem in the past few weeks by various journals (this one included)?

This is what happened. A week ago, *Science* told journalists that it was "lifting the embargo" on two articles concerned with myotonic dystrophy due to be published in its issue of 6 March, roughly two weeks later. Like other journals, *Science* looks askance at premature references to what it is about to publish. So why lift that constraint? Because, it emerged, *Cell* was due to publish last Friday (21 February) an even fuller account of the origin of myotonic dystrophy (Brook *et al.* *Cell* 68, 799; 1972) than the two articles in *Science*'s pipeline. What neither journal knows is that *Nature* then decided not immediately to proceed with the publication of an article provisionally scheduled for 5 March so that its authors could pay attention to some perceptive comments on it.

Happenings of this kind, increasingly commonplace, are demeaning for all concerned — not just for authors, but for the journals in which they seek to publish. Everybody will agree that, in getting to grips with the understanding of an important disease, rapid publication is important; otherwise, physicians will not know what is going on. But how quick is rapid, or can a few days matter all that much? Especially when even the precipitating cause of last week's melodrama, Brook *et al.*, amounts

to an excellent sharpening (but not a resolution of) the questions people have been asking?

The timing of the articles concerned is interesting. The three articles published in *Nature* on 6 February arrived last year on 2 December (Harley *et al.*), 4 December (Buxton *et al.*) and 19 December (Aslanidis *et al.*). The two articles due to appear in *Science* on 6 March were received on 21 January (Fu *et al.*) and 31 January (Mahadevan *et al.*) respectively. *Cell* says that the article by Brook *et al.* was received on 5 February, a mere 16 days before publication. At this rate, bystanders will suppose, delays for research articles on myotonic dystrophy will be down to zero a few weeks from now.

Each of these articles has (or will have) conveyed important news. The two articles to be published in *Science* show that the lengthening repetitive DNA is a replicating nucleotide triplet (in this case, CTG) and that the product of the myotonic dystrophy gene is probably a protein kinase, an enzyme likely to be involved in the tissue-specific regulation of cell activity. Brook *et al.* go further in an interesting way, by confirming what others have concluded and then disentangling the degree to which CTG triplets have expanded on the two separate chromosomes-19s in each cell nucleus.

The way in which these research articles tread on each others' heels must engender as much confusion as enlightenment in the minds of readers. Are we heading for a state of affairs in which what tends to be published is not the 'minimum publishable unit', a slice of salami, but a more substantial record of discovery with something novel tagged onto the end? Those concerned cannot be accused of duplicate publication, because the overlapping parts have not yet appeared in print.

Confusion is further aggravated by the overlapping membership of the author groups. Three of the signatories of Fu *et al.* are also signatories of both Mahadevan *et al.* (in *Science*) and Aslanidis *et al.* (in *Nature*), suggesting that apparently competing groups knew in advance of the overlapping parts. There is nothing wrong with that either, except the hint of a suspicion that people out to maximize the publicity attending discovery have taken to playing tricks on journals. There are two remedies: journals must watch out, especially for the quality of what they publish, and must be uniformly faster.

John Maddox

Wrinkles on the inside

Thorne Lay

THE Earth is a layered planet, with abrupt changes in density and other material properties defining several internal boundaries at various depths. These boundaries are expected to have undulations associated with chemical and dynamical processes in the interior. Shearer and Masters, on page 791 of this issue¹, use seismic waves to obtain the first global map of topographic variations on the mantle discontinuity near a depth of 660 km, the boundary between the upper and lower mantles. Their results are particularly significant given the controversy over the form of mantle convection, which may take in the whole mantle as a single layer system, may be split between separate upper- and lower-mantle circulation systems, or may involve a 'leaky' intermediate regime of penetrative convection².

The primary stratification of the Earth is associated with chemical differentiation which took place when the planet formed and which has developed since. This has resulted in compositional layering, with large chemical differences between the crust, mantle and core. It is generally accepted that the boundaries between these layers — the crust-mantle 'Moho' discontinuity at a depth of a few tens of kilometres and the core-mantle boundary 2,900 km below the Earth's surface — are compositional boundaries across which transport of material is inhibited by strong intrinsic density contrasts. Over the past 25 years various seismological analyses have revealed additional abrupt changes in density,

compressibility and rigidity in the mantle, near depths of 60, 80, 220, 330, 410, 520, 660, 710, 900, 1,050 and 2,640 km (Fig. 1). Some structures, such as the 220- and 330-km features, seem to be intermittent, existing only in certain locations, but the boundaries near 410 and 660 km have the largest contrasts and are accepted as global. Are these also the result of chemical stratification?

Phase transformations of common mantle materials such as olivine are expected to occur near depths of 410 and 660 km. High-pressure experiments show that most of the observed changes in material properties near these depths can be accounted for by the phase transformations with no change in bulk chemistry. Nevertheless, this agreement alone does not preclude the existence of a compositional change in addition to the phase change at these depths. The issue is important, because a compositional contrast across the 660-km discontinuity would probably involve a density contrast sufficient to prevent whole-mantle convection, whereas a purely phase-transformation boundary is more likely to allow significant mass transport between the upper and lower mantle. One observable that can potentially discriminate between a compositional boundary and a phase boundary at 660 km is the magnitude of the topographic relief found at the boundary.

Mapping the topography of the main internal boundaries has long been an objective in geophysics. (Fig. 1). Seismological and gravitational studies demonstrate that crustal thickness (the depth to the Moho) varies from 10 to 70 km, owing primarily to petrological differences in the surface rocks. And there may be smaller undulations on the core-mantle boundary^{3,4}, with depth variations probably being in the range 1–10 km. For the viscous materials that make the Earth, topography on a chemical boundary across which there is a strong density contrast can be sustained only if there is dynamic flow in the adjacent layers. Downwelling regions of the upper-mantle convective system involve cold subducting slabs of oceanic lithosphere, which should substantially depress any chemical boundaries in their vicinity, by as much as

100–200 km (ref. 5). The endothermic phase transition expected for olivine material near 660 km depth is weakly temperature dependent, and if the boundary simply represents a phase transition, it should be depressed by only 20–30 km by the low temperatures of the downwelling slab. Thus, the topographic model in this issue provides a direct test of the fundamental nature of the 660-km boundary, which in turn bears upon the

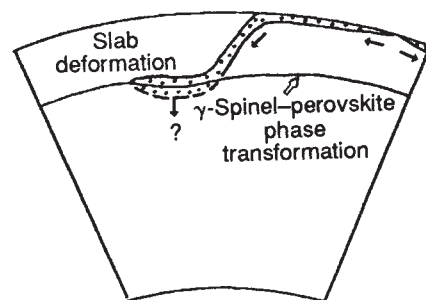


FIG. 2 Schematic interpretation of the cause of variations in the 660-km boundary detected by Shearer and Masters beneath subduction zones. The phase transition from spinel to perovskite, thought to explain the boundary, is deflected downwards as cold material in the subducting slab goes through the transformation. It is not resolved how far the slab material ultimately penetrates.

dynamic configuration of the mantle.

Using a clever combination of seismic waves that provide localized sampling of the 660-km seismic-velocity discontinuity on a global scale, Shearer and Masters find that the boundary depth varies by around 30 km, with regions of upper-mantle downwellings tending to be associated with broad depressions of the boundary. Their procedure is straightforward, but dependent upon our knowledge of the heterogeneous velocity structure in the upper 600 km of the mantle. Fortunately, seismic tomography procedures have placed good constraints on the large-scale velocity structure in the upper mantle over the past eight years^{6–8}. Thus, Shearer and Masters' first-generation results may undergo refinement, but not drastic revision.

The modest depth variations found globally at the 660-km boundary are consistent with results obtained previously from spatially limited sampling⁹. The main implication is that the seismic velocity discontinuity near 660 km is primarily the result of the dissociative phase transition of olivine to perovskite plus magnesiowüstite. The topographic variations in the phase boundary provide an internal thermometer, allowing lateral temperature variations to be mapped. In addition, the broad depression of this boundary in a landward direction from subduction zones indicates cooling of the mantle due to flattening out of the subconducting slab near the boundary (Fig. 2). The question then arises as to why the slab would flatten near the

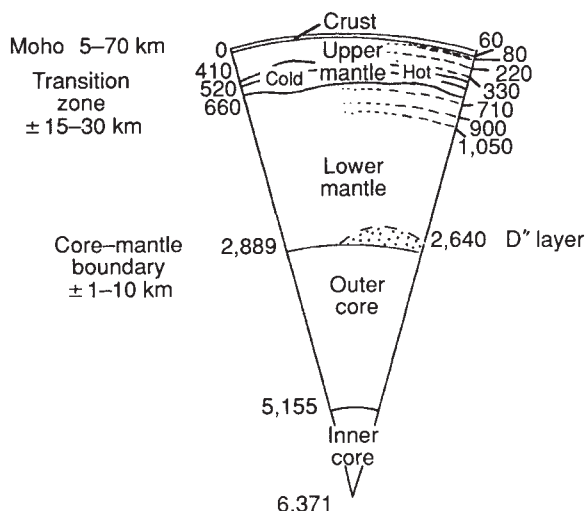


FIG. 1 Schematic cross-section through the Earth identifying major boundaries in the planet, that are observed globally (left) and intermittently (right). Estimates for the variation in depth of each are given where available. The 410- and 660-km boundaries seem to be deflected in the opposite sense by temperature variations in the transition zone, with peak-to-peak variations as large as 60 km possible.

660-km boundary. An increase in viscosity across the boundary associated with the transition could produce strong slab distortion¹⁰. And the precise pressure-temperature relation of the phase transformation could conspire to produce slab deflection, impeding any penetration of the lower mantle¹¹.

These results seem to dispense with models of a chemically layered mantle in which the 660-km discontinuity is viewed as only a chemical contrast which prevents mass transport across the boundary. However, this is not to say that layered convection and a difference in the bulk chemistry between the upper and lower mantles are ruled out. There is continuing debate over whether the perovskite phase transformation near 660 km predicts the correct density for the lower mantle; some evidence indicates that there is a greater proportion here of iron or other high-density material relative to the upper mantle¹². It is plausible that there is a chemical contrast between the upper and lower mantles, associated with deeper seismic discontinuities (at 710 km, for example) or linked to the phase transition near 660 km. Such a chemical boundary may

reflect seismic waves only weakly, so that any strong deflections of the contrast need not be apparent from the seismic studies. Such a chemical contrast could contribute to an increase in viscosity across the mantle transition zone, abetting the deflection of the downwelling slab. Resolution of the overarching issue of dynamic configuration of the mantle is thus not yet in hand, but the work of Shearer and Masters has certainly narrowed some of the options. □

Thorne Lay is in the Earth Science Department, University of California, Santa Cruz, California 95064, USA.

1. Shearer, P. M. & Masters, T. G. *Nature* **355**, 791–796 (1992).
2. Olson, P., Silver, P. G. & Carlson, R. W. *Nature* **344**, 209–215 (1990).
3. Morelli, A. & Dziewonski, A. *Nature* **325**, 678 (1987).
4. Doornbos, D. J. & Hilton, T. J. *geophys. Res.* **94**, 15741–15751 (1989).
5. Hager, B. H. & Raefsky, A. *EOS* **62**, 1074 (1981).
6. Woodhouse, J. H. & Dziewonski, A. M. *J. geophys. Res.* **89**, 5953–5986 (1984).
7. Su, W. & Dziewonski, A. *Nature* **352**, 121–126 (1991).
8. Woodward, R. L. & Masters, T. G. *J. geophys. Res.* **96**, 6351–6377 (1991).
9. Revenaugh, J. & Jordan, T. H. *J. geophys. Res.* **96**, 19763–19780 (1991).
10. Gurnis, M. & Hager, B. *Nature* **335**, 317–321 (1988).
11. Machetel, P. & Weber, P. *Nature* **350**, 55–57 (1991).
12. Jeanloz, R. & Knittle, E. *Phil. Trans. R. Soc. A* **328**, 377–389 (1989).

LASER PHYSICS

Ultrason et lumière

R. G. Evans

RECENT developments in micro- and nanoengineering have been brought about only through the application of the most sophisticated electron and atomic beam technology. But now K. R. Chen and J. M. Dawson propose what must be the simplest example of micro-engineering yet discussed (*Phys. Rev. Lett.* **68**, 29–32; 1992), with the promise of tunable coherent radiation in any part of the spectrum, from microwaves to soft X-rays, from a single device. To make the periodic structure that forms the heart of their proposed laser, Chen and Dawson take the most straightforward approach: using ultrasonic sound waves.

The best-known tunable source of radiation in the optical region is the free electron laser (FEL). This works by modulating the motion of high-energy electrons using the periodic magnetic field produced by a 'wiggler' magnet. The undulating field induces the electrons to radiate and, as the system is periodic, the radiation reacts back on the beam electrons to bunch them together. This gives an exponential growth in the intensity of scattered radiation. The electron beam moves relativistically, so that the relatively long wavelength of the wiggler magnet (say 1 cm, corresponding to radio waves) is Doppler shifted to far shorter wavelengths. The radiation

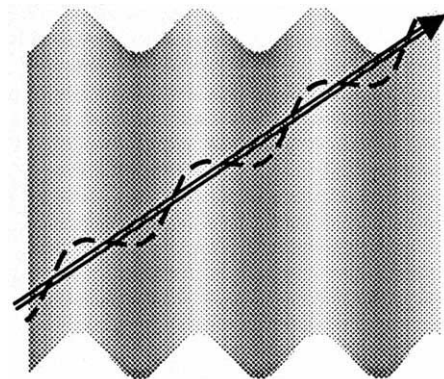
frequency is tuned through the optical spectrum by changing the beam energy.

The drawbacks of the FEL are many. In particular, producing high-frequency radiation requires high-energy electrons, in the range 10–100 megaelectronvolts. And small, powerful, precise and, consequently, expensive wiggler magnets are necessary. Chen and Dawson's new contribution is to note that the periodic modulation of the electron beam can be achieved equally well by a density variation in a plasma of ions and electrons, and that the easiest way to make such a periodic variation is with a sound wave. Normally the multitude of instabilities that plague plasmas would undo any regular density modulation, but Chen and Dawson propose using a method developed at Imperial College to make a relatively cold plasma, immune to the instabilities, using multiphoton ionization. The density of the plasma is exactly equal to the background density of the gas before ionization.

Chen and Dawson's ion-ripple laser is delightfully simple: set up a standing ultrasonic wave in a hydrogen gas cell, ionize the gas with a short, intense pulse of laser light, and fire through the resulting plasma an electron beam of just a few megaelectronvolts energy. Because of the very short wavelength of the

sound wave, compared with that of a wiggler magnet, only a moderate Doppler shift is needed to boost the electrons' radiation to optical frequencies and the efficiency can be much higher than with the 'mechanical' FEL.

The ion-ripple laser has the potential to be more of a table-top device than the FEL, through its reduced requirement



The path of a high-energy beam of electrons, passing through a plasma excited by a standard ultrasonic wave, will deviate about its mean. Laser radiation will emerge along the beam direction.

for high-energy electrons. It is also delightfully simple to tune; of course, it is not that difficult to change the energy of the electron beam in an FEL, but it is more elegant to tune your laser radiation by altering the pitch of a sound wave.

At first sight, the ion-ripple laser has many features in common with the channelling radiation produced by electron beams passing through ion arrays in crystals. But the three-dimensional structure of solids and the higher background electron densities mean that channelling radiation is a less controllable phenomenon.

There are diverse applications possible for a table-top tunable radiation source: selective excitation and ionization make possible new measurements in atomic physics; selecting vibration frequencies in molecules can produce specific radicals and initiate chemical reactions; colour selectivity can be biologically and medically useful; and narrow-band tunable radiation can be used to investigate the surface properties of new engineering materials such as low-dimensional structures and quantum wells.

On a more speculative level, our humble piece of microengineering, using nothing more sophisticated than a sound wave, may well find application in the preparation of new generations of microelectronic circuits. Yet again we may see the applications of science coming full circle to feed the next generation of sound and sometimes brilliant ideas. □

R. G. Evans is at the Atlas Centre, Rutherford Appleton Laboratory, Chilton, Oxford OX11 0QH, UK.

Biology's wiring circuits

R. J. P. Williams

ONE of the salient problems in understanding the energy-capture machinery of biology, especially in photosynthesis and in mitochondria, is electron transfer over a long distance, R , up to 20 Å. The topic is taken up on page 796 of this issue¹ by C. C. Moser and colleagues, who describe a detailed study of one particular electron transfer chain, and a comparative analysis of others.

Appreciation of the value of the study must start with an appreciation of the problem. In both mitochondria and thylakoid membranes, and in the corresponding apparatuses in bacteria, there are chains of organic cofactors, chlorins and quinones for example, and chains of metal ions, including iron, copper, manganese and haem, which are arranged within membranes so as to assist electron flow and generate charge separation, the beginning of energy capture. Even in many soluble redox enzyme particles, shorter chains of these electron-transfer centres are found and many models have been synthesized (see Fig. 2 of the paper, page 799). The general question then arises as to how an electron can move from one centre to the next, 10–20 Å at a time, in a systematic and ordered fashion.

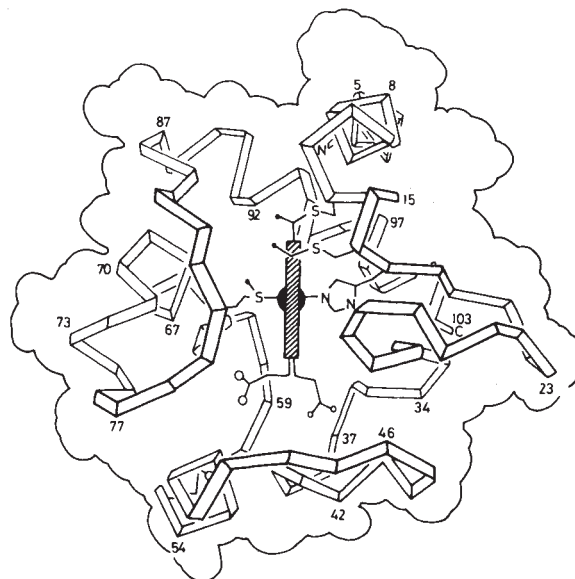
Part of the answer lies in the placing of the oxidation/reduction centres in the order of their potentials expressed in terms of free-energy differences, ΔG , so that flow is often downhill in a thermodynamic sense. This is not of concern here. Another part of the answer lies in the placing of individual centres in an eccentric fashion within a protein, so that the centre is much closer to one part of the protein surface than all others (see the figure accompanying this article). In this way use is made of the dependence of electron transfer probability on distance, because only in certain directions is R less than 20 Å. It is this distance dependence that Moser *et al.* tackle. The chain of reaction centres is then determined by the organization of proteins, with little risk of electrons leaking into incorrect pathways.

The factors controlling electron transfer were put into a useful theoretical framework by Marcus², and include one other term, the reorganization energy, λ , necessary to equalize bond lengths of the oxidized and reduced states of a reaction centre if electron transfer is to occur, which, together with the ΔG , must be corrected for if an analysis of distance dependence is to be made. There have been similar approaches over the intervening years (as shown in the boxed material on page 798), but Moser *et al.*

have looked at results in terms of the Marcus expression for electron transfer rate, k_{et}

$$k_{et} = \frac{2\pi}{\hbar} V_0^2 e^{-\beta R} \{FC\}$$

where R is the distance between centres; FC , including λ and ΔG , is the Franck-Condon term to be taken into account before distance dependence can be



View of a typical electron-transfer protein, cytochrome *c*, where (1) the metal ion centre is buried some 5–10 Å inside the protein in one direction (here from the top surface of the protein, that is from above the plane of the page), but in all other directions is more than 10 Å from the surface (the protein in projection is roughly a circle of 30 Å diameter); (2) the metal centre cannot exchange ligands and the protein does not undergo large conformational changes; and (3) the centre is screened from water. The iron is shown as a black ball and the porphyrin as a disk. The protein chain folds into a homogeneous medium for electron transfer, as discussed by Moser *et al.*¹

evaluated; β is a coefficient of distance coupling; and $(2\pi/\hbar) V_0^2$ is the maximum coupling at $R = 0$ (that is, closest contact between centres).

The system chosen for reference is the bacterial photoreaction centre, which has great advantages in that the distances between centres are known from the crystal structure; none of the centres can rearrange their solvent spheres (that is, they are not exposed to water); and the proteins do not undergo rearrangement during reaction because they are not coupled to events such as proton release. Given that there are many centres in the chain, and many possible pair-wise electron-transfer reactions that can be measured, the system provides a definitive examination of the equation.

The expectation from such analysis, once corrections have been made for the

other terms influencing the rate constant, is that there should be a straight-line relationship between $\log k_{et}$ and the distance R . The intercept at $R = 0$, that is for close contact of reacting electron transfer sites, is $(2\pi/\hbar) V_0^2$ and is expected to be of the order of a vibration frequency ($k_{et} = 10^{13} \text{ s}^{-1}$). Moser *et al.* find just such an intercept. The slope of the line, the second finding, is $\beta = 1.4 \text{ Å}^{-1}$, which implies that there will be insignificant electron transfer rates at 20 Å separation of centres. The construction of all known pairs of electron-transfer centres in biology indicates that

the eccentric placing of the centres (see Figs 1 and 2a, pages 797 and 799) allows an approach at or below 20 Å so that this experimental value of β is highly satisfying. In model systems, as Moser *et al.* show, the value of β may be different, but this may be just a global 'solvent' effect. In some publications $\beta = 0.9 \text{ Å}^{-1}$ has appeared for protein systems³, but this value must now be subject to reexamination, especially because one of the electron transfer centres is in water.

A further striking feature of the new results is that the authors have to refer only to the known electron-transfer sites. Thus the protein does no more than provide an averaged medium, and a change to another medium is likely to affect the rates of electron transfer (as Moser *et al.* point out). There is no suggestion here that the electron passes through a series of particular bonds or makes a significant differential use of protein side-chains. A feature of the experimental material is, however, that all the centres have low absolute redox potentials (less than +0.5 volts).

In other chains, such as that of photosystem II, the redox potentials exceed +0.8 volts. Here, and for example in some particulate redox enzymes (and in some models), and in the cytochrome oxidase protein, it is much more likely that the aromatic side-chains of proteins in particular, but note also the peptide bonds, enter into the electron-transfer

1. Moser, C. C., Keske, J. M., Warncke, K., Farid, R. S. & Dutton, P. L. *Nature* **355**, 796–802 (1992).
2. Marcus, R. A. & Sutin, N. *Biochim. biophys. Acta* **811**, 265–322 (1985).
3. Axup, A. W., Albin, M., Mayo, S. L., Crutchley, R. J. & Gray, H. B. *J. Am. chem. Soc.* **110**, 435–439 (1988).
4. Williams, R. J. P. *Molec. Phys.* **68**, 1–23 (1989).
5. Wikström, M. & Babcock, G. T. *Nature* (in the press).

process. In fact in photosystem II, and in some enzymes, specific tyrosine and tryptophan radical redox centres are known to be part of an electron-transfer specified route or chain⁴.

If this work is generally accepted then the next problem associated with electron transfer in biological systems is the coupling within membranes to proton

transfer pumps. So the work of Moser *et al.*¹, together with knowledge of proton flow⁵, will reveal the nature of biology's wiring circuits and local energizations associated with energy transduction. □

R. J. P. Williams is in the Inorganic Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QR, UK.

ACTIVE GALAXIES

Invisible jet sheds new light

Simon Garrington

AT radio wavelengths, the sky is dominated by radio galaxies and quasars whose enormous radio luminosities (10^{36} watts) come from twin lobes straddling their host galaxies. Detailed radio images produced over the past 15 years by aperture synthesis telescopes, such as the Very Large Array (VLA) in New Mexico and MERLIN in the United Kingdom, have provided striking confirmation that the lobes are fed by narrow jets of plasma created by the active nucleus of the galaxy. But although the lobes are usually symmetric about the nucleus, the jets — especially in the more powerful sources — are often one-sided. One of the most famous one-sided jets, that in M87, was in fact first detected optically over 70 years ago. The papers by Stiavelli *et al.*¹ and Sparks *et al.*² in this issue use new optical images to show that there must be an unseen counterjet in M87, prompting a re-examination of what we do see in radio jets.

Three explanations have been advanced for the asymmetry of the jets in

radio galaxies and quasars: that they are truly one-sided but flip from side to side; that they are two-sided but only on one side is there enough energy dissipated to reveal the jet; and that the two jets are moving at speeds close to the speed of light. In the last case, the emission is strongly beamed along the direction of motion by relativistic aberration: the approaching jet will then appear much brighter than the receding one.

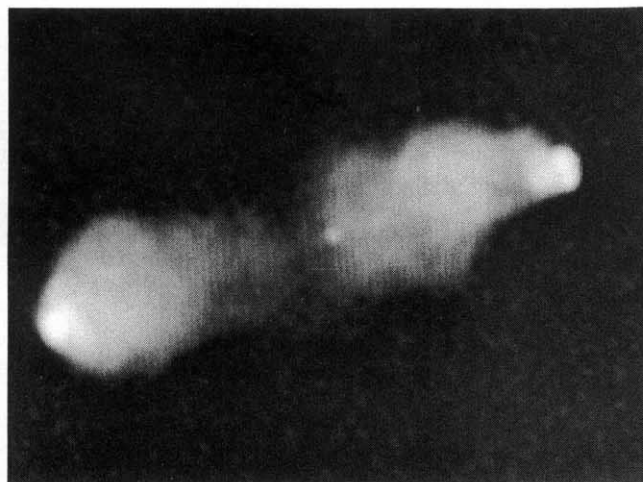
For quasars, the beaming model has received strong support from observations using radio arrays spanning Europe and the United States (VLBI), which have sufficient angular resolution to track the motion of features in the inner jets over periods of a few years. The apparent speeds exceed the speed of light, which is naturally explained by special relativity if the velocity is close to the line of sight. Such superluminal motions have now been followed out to several hundred light-years³ and are sufficient to account for the jet asymmetry (see also ref. 4).

The situation is less clear with radio galaxies. The powerful FR2-type radio galaxies (see box) have jets similar to those in quasars — results so far suggest slower but nevertheless relativistic speeds⁵. This supports unification schemes according to which the FR2 galaxies are simply quasars at larger angles to the line of sight⁶. But there is much less agreement on the speeds of the jets in the less-powerful FR1 galaxies such as M87. Numerical simulations⁷ and theoretical models⁸ suggest trans-sonic velocities ($1,000\text{--}10,000\text{ km s}^{-1}$), well short of the speed of light, c . In M87 itself, the brightest features in the jet have measured speeds of only $0.3c$, too slow to produce strong relativistic beaming⁹. Because there is no hint of a counterjet in M87, the implication is that this is a truly one-sided jet.

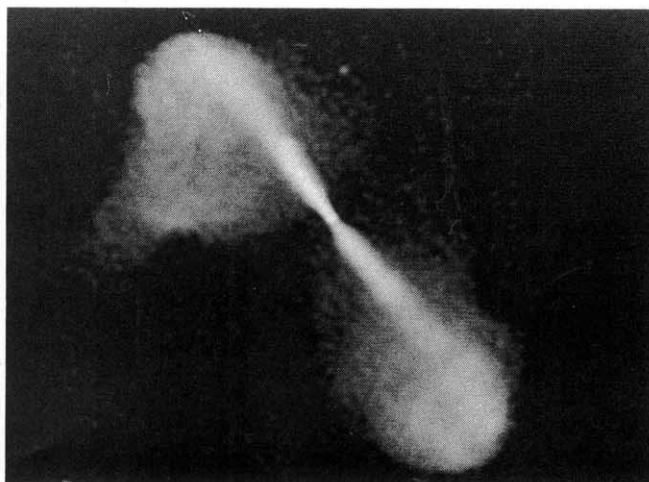
The observations reported on pages 802 and 804, respectively, by Stiavelli *et al.*¹ and Sparks *et al.*² show for the first time that there must be an unseen counterjet in M87. The crucial new evidence comes from the identification of an optical counterpart to the radio hotspot opposite the visible jet. By measuring its polarization Sparks *et al.* confirm that the optical emission is, like the radio emission, synchrotron radiation produced by electrons spiralling around magnetic field lines. Because the radiative lifetime of the electrons varies with the emitting frequency ν as $\nu^{-1/2}$, the optical data set a much lower limit on the age of this hotspot than did previous radio images and so rule out the possibility that it is a relic of a defunct jet.

So we must accept either that the counterjet propagates so efficiently that

Two types of extragalactic jets



EXTRAGALACTIC radio jets come in two flavours. Powerful radio galaxies (FR2 sources), such as 4C14.11 shown on the left, have weak, narrow jets which end in bright hotspots at the edges of the radio lobes. Numerical simulations suggest that these jets are light and supersonic. Less-powerful radio galaxies (FR1 sources), such as 3C296 shown on the right, have brighter, broader jets



which do not end in hotspots but fade into diffuse plumes. These jets are often thought to be turbulent and trans-sonic. The narrow FR2 jets are almost always one-sided, whereas the broader FR1 jets are often asymmetric and very occasionally, as in M87, completely one-sided. (Figures are VLA images, courtesy of J. P. Leahy.) S.G.

it leaves no trace or that relativistic effects are in fact important in FR1 jets. Both options challenge our conventional assumptions about extragalactic jets and raise the question: just what do we see in the visible jet? On evidence from radio images¹⁰, Sparks *et al.* suggest that the visible jet is revealed only by some form of surface excitation and that this mechanism is simply not triggered in the counterjet. If so, how many other jets could be hidden in this way?

Alternatively, it may be the apparent velocity that has deceived us: Stievelli *et al.* point out that the most recent determination of motion in the outer jet reveals significantly higher velocities — up to $(1.0 \pm 0.3)c$ — which are not constant along the jet¹¹. One interpretation is that the jet has a speed close to c and the features we see are slow-moving patterns in the flow. Because it is the flow speed, not the pattern speed, that determines the relativistic beaming, this could easily account for the jet asymmetry. This challenges the assumption that FR1 jets are slow, but supports the idea that relativistic beaming in the nucleus

could make an FR1 galaxy appear as a BL-Lac-type active galaxy if seen at a smaller angle to the line of sight¹². But why do the superluminal jets apparently move with the full flow speed? □

Simon Garrington is at the Nuffield Radio Astronomy Laboratories, Jodrell Bank, Macclesfield, Cheshire SK11 9DL, UK.

1. Stievelli, M., Biretta, J., Moller, P. & Zeilinger, W. W. *Nature* **355**, 802–804 (1992).
2. Sparks, W. B., Fraix-Burnet, D. & Machetto, F. *Nature* **355**, 804–806 (1992).
3. Davis, R. J., Unwin, S. C. & Muxlow, T. W. B. *Nature* **354**, 374–376 (1991).
4. Heckman, T. & O'Dea, C. *Nature* **354**, 355 (1991).
5. Preuss, E. *et al.* in *Proc. NRAO Workshop on Parsec-scale Radio Jets* (eds Zensus, J. A. & Pearson, T. J.) 120–124 (Cambridge University Press, 1990).
6. Barthel, P. D. *Astrophys. J.* **336**, 606–611 (1989).
7. Norman, M. L., Winkler, K.-H. A. & Smarr, L. in *Physics of Energy Transport in Extragalactic Radio Sources* (eds Bridle, A. H. & Eilek, J. A.) 150–167 (NRAO, Greenbank, 1984).
8. Bicknell, G. V. *Astrophys. J.* **305**, 109–130 (1986).
9. Biretta, J. A., Owen, F. N. & Cornwell, T. J. *Astrophys. J.* **342**, 128–134 (1989).
10. Owen, F. N., Hardee, P. E. & Cornwell, T. J. *Astrophys. J.* **340**, 698–707 (1989).
11. Biretta, J. A. & Owen, F. N. in *Proc. NRAO Workshop on Parsec-scale Radio Jets* (eds Zensus, J. A. & Pearson, T. J.) 126–128 (Cambridge University Press, 1990).
12. Browne, I. W. A. *Mon. Not. R. astr. Soc.* **204**, 23p–27p (1983).

EVOLUTIONARY BIOLOGY

Byte-sized evolution

John Maynard Smith

A NEW way in which biologists may learn something from computers is suggested by a paper by T. S. Ray¹, in which he describes the behaviour of an open-ended evolving system.

Ray's approach differs from the more familiar ones of genetic simulation and genetic algorithms. We have learnt a lot from simulation of genetic systems, but the method does not differ in principle from the simulation of physical or economic systems. Genetic algorithms, pioneered by Holland², are nearer to Ray's approach. They attempt to solve practical problems — the design of an aeroplane's wing or a transport system — by using an analogue of a genetic system, with mutation and selection, to search for an optimal solution. They are of interest to evolutionary biologists in telling us how effective particular features — for example, recombination — are in mediating evolution. But they are not open-ended, because the 'fitness' of particular sets of instructions, and the ultimate objective, are set by the programmer. Hence they can tell us little about how populations of evolving organisms will behave. Ray has developed an open-ended evolving system, essentially by making the fitness of the entities depend only on their ability to survive and replicate in the computer world he has created.

His starting point is a set of 80 instruc-

tions which, placed in the memory of a computer and depending on its operating system, causes copies of itself to appear in the memory. Because he wrote the instructions himself, his model is not relevant to the origin of life, but only to its further evolution. If there were no deaths and no mutations, the memory would fill up with copies and the process would then end. Random mutations occur when instructions are copied, and, at a lower rate, between replications. Deaths occur because new individuals are placed at the bottom of a 'queue', and die when they reach the top: the population is allowed to fill about 80 per cent of the memory.

The computer allocates a small slice of time in turn to the creatures in the memory soup: during that time they may replicate once, or several times, or not at all. Daughter creatures are placed next door to their mothers.

The evolutionary behaviour of the system is surprisingly rich. The first important variants to arise are 'parasites', 45 instructions long, which cannot replicate on their own, but use the copying procedure of a neighbour: they are an exact analogue of satellite viruses. Once parasites become common, 'hosts' may evolve immunity; then new types of parasite evolve that are able to attack the new hosts. Evolutionary arms races of this kind have often been postulated

and occasionally observed. Next there arise 'hyperparasites' which, by an ingenious trick, persuade the parasites to replicate them. These hyperparasites drive the parasites to extinction, because the latter are no longer replicating themselves. Although a single hyperparasite could not replicate itself, and so would die from the accumulation of mutations, a group of neighbours are able to do so. In Ray's terminology, they are social hyperparasites. It is tempting to draw an analogy with coviruses, but this is not exact, because cooperating hyperparasites are identical, whereas coviruses are different and complement one another's inadequacies. But it is interesting that conditions are ideal for the evolution of cooperation by kin selection, because daughters live next door to their mothers.

More complex, and still largely unanalysed, changes occur in longer runs. Communities dominated by individuals with 400–800 instructions arise. Absolute stasis seems not to happen, but a community may remain comparatively constant in character for a time, before changing rapidly and chaotically to a community with a different distribution of size classes. Extinction of an entire community was observed twice. In both cases it occurred when the dominant size class contained a number of instructions about equal to the reciprocal of the mutation rate; it had hit what Eigen³ has called the error threshold, although I have previously thought of the threshold as imposing an upper limit on the genome size of individuals, rather than as causing the extinction of whole communities.

Ray suggests that his creatures had to have three features, all copied from real biology, if evolution was to occur. First, his programming language, Tierran, had to have a small instruction set — in fact, 32 instructions — analogous to the small number of 64 codons. This is related to the requirement that mutations should have a reasonable chance of making functional sense: similarly, biological evolution can happen only if a reasonable proportion of amino-acid sequences make functional proteins, at least locally in protein space⁴. Second, the Tierran language, unlike most computer languages, does not specify an address by an integer, but by template matching. For example, the 'jump' instruction is followed by a sequence of four bits: the address is then the nearest complementary sequence of four bits. This is analogous to the fact that a protein molecule does not find its point of action by moving to a place specified by coordinates, but by diffusing until it finds a complementary surface to which it can bind. Third, the creatures are 'cellular', in the sense that no creature can alter, or

write over, the instructions of another. Interactions between creatures occur because one can make use of the instructions of another, just as a virus can make use of the gene products of the cell, or of other viruses. It is interesting, however, that there is no distinction between genotype and phenotype: in this it resembles an RNA world.

What can such a study tell us about biology? There is a danger that, in Boyd and Richerson's words⁵, we shall replace a world we do not understand by a model of the world we do not understand. Ray could reply that one can do experiments in his world more easily than in the real one. And we badly need a comparative biology. So far, we have been able to study only one evolving

system and we cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we will have to look at artificial ones. Ray's study is a good start. □

John Maynard Smith is in the School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK.

1. Ray, T. S. in *Artificial Life II*, Santa Fe Inst. Studies in the Sciences of Complexity, Vol. XI (eds Langton, C. G., Taylor, C., Farmer, J. D. & Rasmussen, S.) 371–408 (Addison-Wesley, Redwood City, California, 1991).
2. Holland, J. H. *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, 1975).
3. Eigen, M. *Naturwissenschaften* **58**, 465–523 (1971).
4. Maynard Smith, J. *Nature* **225**, 563–564 (1970).
5. Boyd, R. & Richerson, P. J. *Culture and the Evolutionary Process* (University of Chicago Press, 1985).

OZONE DEPLETION

Cold comfort in the greenhouse

Jeffrey T. Kiehl

MAN-MADE chlorofluorocarbons (CFCs) are important not only for the part they play in ozone destruction. They are also greenhouse gases, letting through short-wave solar radiation, but trapping long-wave radiation from the Earth's surface and lower atmosphere. So it may be with a false sense of relief that readers learn on page 810 of this issue¹ that the reduction of greenhouse forcing that follows from ozone destruction (ozone is another greenhouse gas) almost offsets the increase from the presence of CFCs. But put in this globally averaged way, the result ignores the complexities that dog the climate system.

Concerns with the environmental effects of CFCs were first raised in the mid-1970s, when it was suggested² that man-made chemicals such as CFCs could efficiently destroy ozone in the upper stratosphere, 12–50 km above the Earth's surface. The modelled effects of increased CFCs also indicated an associated increase in lower-stratospheric ozone, due to what was termed the self-healing process³: the depletion at higher altitudes allows solar radiation to penetrate further into the stratosphere, promoting ozone-forming photochemical reactions.

Since this early work, the study of stratospheric ozone depletion has been punctuated by theoretical and observational surprises. Foremost among these was the discovery, in 1985, of the Antarctic ozone hole⁴. This depletion, which is linked to the presence of ice clouds in the lower polar stratosphere, was not predicted in any chemical model. The most recent observations indicate that over the past decade (1979–1990) significant depletions have occurred at mid- and high latitudes in the lower

stratosphere⁵, in contradiction to the self-healing process.

In addition to the chemical importance of ozone, there is the significant radiative role of this atmospheric constituent, addressed by Ramaswamy *et al.* in this issue¹. The authors compare the climate impact of several greenhouse gases, in particular CFCs, with and without the radiative effects of lower stratospheric ozone depletion.

It was shown in 1975 that CFCs have important implications for the greenhouse warming effect⁶. This is owing to their strong longwave absorption strength and the rapid increase in their abundance in the atmosphere. Ramaswamy *et al.* show that the reduced ozone level in the lower stratosphere results in a negative forcing of the climate system that, in the global average, essentially cancels the direct positive greenhouse effect of the CFCs. Thus, the net globally averaged effect of CFCs (direct greenhouse plus indirect chemical) on the potential greenhouse warming problem is quite small.

These findings result from the interesting radiative role of ozone in the lower stratosphere. In this region of the stratosphere, ozone tends to warm the atmosphere locally through both short-wave and longwave radiative processes. With the addition of CFCs and the concomitant depletion of ozone, three important changes occur. First, there is less ozone available to absorb solar radiation and warm this part of the atmosphere. This leads to a cooling of the lower stratosphere. Second, the reduction in ozone in the lower stratosphere allows more solar radiation to penetrate into the lower atmosphere, or troposphere, which warms the Earth's

surface. This effect actually increases the warming potential of the CFCs. Third, the reduction of ozone and the cooler lower stratosphere results in less emission of longwave radiation down into the troposphere. This leads to a cooling of the surface–troposphere system. It is the third process that results in the cancellation of the direct greenhouse effect of the CFCs.

But these findings do not imply that the increasing quantity of CFCs in the atmosphere is not a problem. Although there is a global cancellation of the CFC warming potential, the magnitude of the negative forcing depends upon latitude, owing to the latitudinal variation in ozone depletion. The negative forcing, then, is largest at mid- and high latitudes, and weakest at the Equator. The latitudinal distribution of the total greenhouse warming effect due to CO₂, CH₄, N₂O and CFCs thus decreases more sharply at mid- to high latitudes when the indirect ozone effect is included. The predicted greenhouse forcing for the period 1979–90 decreases from a tropical maximum of 0.5 W m⁻² to a polar minimum of 0.3 W m⁻² with no ozone effect. The ozone depletion in fact slightly increases the tropical forcing, but reduces the polar forcing to nearly zero. Thus, the latitudinal gradient of the greenhouse forcing is more than doubled by the inclusion of the lower stratospheric ozone forcing. This strong increase in the gradient of greenhouse forcing could have important dynamic effects. In the present work, Ramaswamy *et al.* use models that fix dynamical feedback effects, and they point out that general circulation models will be needed to address any dynamical issues.

The new results show that stratospheric ozone continues to surprise atmospheric scientists. It would be wrong to conclude from them that CFCs are a benign component in global warming. Indeed, these findings indicate how intimately coupled is the chemistry of the stratosphere to the climate response at the Earth's surface. From the past 20 years of ozone studies, it seems that the only thing that is not surprising about this facet of the climate is that it will keep surprising us. □

Jeffrey T. Kiehl is at the National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado 80307, USA.

1. Ramaswamy, V., Schwarzkopf, M. D. & Shine, K. P. *Nature* **355**, 810–812 (1992).
2. Molina, M. J. & Rowland, F. S. *Nature* **249**, 810–812 (1974).
3. WMO/UNEP Global Ozone and Monitoring Project, Report No. 18 (World Meteorological Organisation, Geneva, 1988).
4. Farman, J. C., Gardiner, B. G. & Shanklin, J. D. *Nature* **315**, 207–210 (1985).
5. WMO/UNEP Scientific Assessment of Stratospheric Ozone Depletion, 1991 (World Meteorological Organisation, Geneva, in the press).
6. Ramanathan, V. *Science* **190**, 50–52 (1975).

No preservatives

If instant coffee keeps a good part of its native virtues when it is lyophilized (freeze-dried), why not also red blood cells? There are various ways of storing material for transfusion — wet, dry, frozen, or even perhaps in the form of cross-linked haemoglobin. Each has its advantages; but if cost is a primary criterion, R. P. Goodrich *et al.* (*Proc. natn. Acad. Sci. U.S.A.* **89**, 967–971; 1992) may have the answer. For lyophilized cells, they find, can be stored at room temperature and their metabolic state is only mildly perturbed by what they have been through — no more so, at all events, than that of conventionally banked cells. It is true that the drying takes seven days of pumping, but the argument is that this will still be cheaper than maintaining refrigeration plants. Soon you may be able to carry sachets of your own dried red cells in your pocket with your car insurance and donor card.

Chemistry set

With all this talk about nanotechnology (see page 761), chemists are keen to restake their natural claim to this domain. In a virtual manifesto (*J. Am. chem. Soc.* **114**, 601–621; 1992), P. Kaszynski, A. C. Friedli and J. Michl describe their ambitions for a molecular “Tinkertoy” set, with which one could make mechanically strong designer solids as light as aerogels. Chemical substitutions would endow the products with special properties. The authors outline the synthesis of rigid molecular rods built up from individual 5-carbon, propellane molecules. They obtain these ‘staffane’ rods in only small yields, but find them to be stable to heat, air and common solvents, and transparent to light. Kaszynski *et al.* concede that synthesizing the spools needed to connect these rods will be the hard part.

Wash out

J. SMIT *et al.* (*Geology* **20**, 99–103; 1992) have come up with support for the idea that the Yucatán area of Mexico was hit by a large extraterrestrial body at the Cretaceous/Tertiary boundary. The authors looked at an outcrop of rock in northeastern Mexico which dates biostratigraphically to around that time, and which was deposited under gentle conditions some 400 m underwater. Running through the outcrop is a distinctive unit topped by a band anomalously rich in iridium and with features consonant with there having been a disruptive event nearby: the lower part contains tektites (glass droplets) and minerals bearing the scars of impact shock; above that lie coarse sediments and plant remains. Smit *et al.* interpret the one as the product of the impact's immediate fall-out, the other as larger deposits from giant waves stirred up in the Gulf of Mexico.

Models of the seismic cycle

Christopher H. Scholz

COULD the Loma Prieta earthquake of 1989 have been predicted from precursory seismic activity in California? Quite possibly, suggest B. E. Shaw, J. M. Carlson and J. S. Langer in the *Journal of Geophysical Research* (**97**, 479–488; 1992), on the basis of computerized simulations of a mechanical model of earthquake activity.

Models of spatially extended, dissipative systems which exhibit the property known as self-organized criticality have been widely discussed as models of earthquakes (see my News and Views article in *Nature* **348**, 197–198; 1990). Their main claim to fame as earthquake models is that they mimic the Gutenberg–Richter law for the frequency of earthquakes of different sizes. There are now quite a few variations of this class of model that can join the club if this is the sole basis for membership. Shaw, Carlson and Langer delve more deeply into their particular model, attempting a much more ambitious comparison of its behaviour with natural seismicity.

Their model is a one-dimensional spring–slider chain (see Fig. 1) in which the block slider motions are solved dynamically. It produces a spectrum of small-scale events that exhibit scaling (frequency reducing with size) and occasional large events which do not belong to this scaling relation. Natural earthquakes also fit into two classes, and J. F. Pacheco, L. R. Sykes and I have recently shown that large and small earthquakes have different size distributions (*Nature* **355**, 71–73; 1992).

Shaw and colleagues compare, for their model, the temporal and spatial relationship of the small events to the large ones, showing that the small-event activity increases and becomes more

concentrated towards the end of the large-event cycle, which might be understood as a form of precursory activity for large events. Their comparison with real earthquake catalogues gives mixed results. In general, the natural seismicity did not show any clear indication of this clustering, and in

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

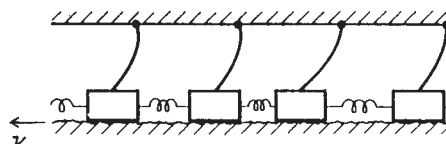


FIG. 1 A chain of sliding blocks connected by springs can behave like a geological fault. (From Shaw *et al.*)

FIG. 2 Cooking dinner in the Santa Cruz Mountains after the Loma Prieta earthquake cut off the power supply.

many cases evolved quite differently.

The one exception was an acceleration of activity on the Calaveras fault, connected to the San Andreas fault, in the years preceding the 1989 Loma Prieta earthquake. Long-term acceleration of activity in broad regions surrounding this and other earthquakes has been noted by L. R. Sykes and S. C. Jaumé (*Nature* **348**, 595–599; 1990). In that study, the authors weighted the data according to earthquake moment, which favours large events, rather than event counts, as used by Shaw *et al.*, which emphasizes smaller events. The convergence of these results, and the guidance provided by the model, suggest that a more diligent examination of seismic catalogues is now in order.

Shaw and colleagues are the first to attempt to apply this type of model to

the behaviour over a whole seismic cycle, in particular addressing the issues more closely related to earthquake prediction that are of broadest public interest. It will come as no surprise to seismologists that the study was not a signal success. If such precursory clustering of activity were common, it would have been noticed long before. But the model is still crude. It should be of little surprise that a homogeneous, one-dimensional model fails to mimic a two-dimensional fault that may be hetero-

geneous at many scales. It may be better to test the model against a simpler system. In this spirit, the observation (see D. A. Lockner *et al.* *Nature* **350**, 39–42; 1990) of spatial and temporal clustering of acoustic emissions in the run up to fracturing of intact rock in laboratory experiments will be an encouragement to Shaw and colleagues. □

Christopher H. Scholz is at the Lamont–Doherty Geological Observatory, Palisades, New York 10964, USA.

MOLECULAR ECOLOGY

Small, green and different

Terry Burke

THE main business of evolutionary and behavioural ecologists is to understand adaptation, and they have plenty of raw data to worry about. Direct experimental approaches are often not possible, and so the comparative method is used to try to infer underlying evolutionary processes^{1,2}. The principal worry in comparative analyses is — or should be — the lack of statistical independence among data points. This is primarily because the organisms being compared may share characteristics simply because they have both inherited them from a common ancestor.

On page 817 of this issue³, Richman and Price neatly take advantage of methodological advances in both molecular systematics and comparative analysis to demonstrate pronounced adaptive differences in habitat selection, feeding behaviour and morphology amongst eight sympatric species of *Phylloscopus* leaf warblers. These species are otherwise rather similar (small and greenish, see Fig. 1), and a popular challenge for the more masochistic bird-watcher; their most obvious difference is their song. The new study is remarkable for its successful integration of quantitative data on morphology, behaviour, ecology and phylogeny. It shows that where behavioural characteristics associated with feeding and habitat choice are plotted against multivariate morphological size variables, they are found, as expected, to be closely correlated, and the more closely related of these species are indeed often rather similar to one another. More revealingly, when the effects of phylogeny are controlled for, it is the differences between species, or groups of species, in behaviour and morphology which are associated in a predictable, adaptive way.

Darwin used the comparative method extensively, and for most of the time since then the method has been essentially unchanged. In an attempt to reduce the extent of nonindependence, the

first statistically based methods used nested analyses of variance to identify the taxonomic level at which most variation was apparent, before proceeding to use this level as the unit of analysis¹. This did not, of course, completely circumvent the problem of shared ancestry. Alternative methods which take the species phylogeny into account have therefore been proposed^{4,5}.

For example, in the case of a qualitative (presence/absence) character, if a phylogeny is available for the taxa under study then it is possible to infer how often the character has evolved and with which variables it is associated (*a* in Fig. 2). This approach was used, for instance, to show that although there is a strong and unsurprising correlation between a bird species having a lek mating system (where mating occurs at a communal display area and males take no part in parental care) and being sexually dimorphic for size and plumage, this correlation disappeared when an out-

group analysis⁵ was used to control for common ancestry⁶.

The snag for practitioners of the comparative method has been that good phylogenies are in short supply. Richman and Price tackled the problem by constructing their own from an impressive 910 base pairs of mitochondrial cytochrome *b* gene sequence from each of 20 species. Obtaining such phylogenetically informative data has suddenly become an almost routine task, thanks to the availability of the polymerase chain reaction technique and of relatively universal oligonucleotide primers for haploid, and suitably rapidly evolving, DNA sequences in animal mitochondria⁷ or plant chloroplasts⁸. Having obtained a phylogeny, Richman and Price completed a comparative analysis in which they were able to control for common ancestry using Felsenstein's method of independent contrasts⁴. This is not the first time Felsenstein's method has been applied⁹, but will probably encourage its more frequent use.

DNA sequences provide phylogenetic data which are, usefully, largely independent of other characters. Many alternative phylogenetic trees may, however, be reasonably compatible with the raw data. Thus a bootstrap analysis was used to examine the robustness of the branching pattern in the *Phylloscopus* tree; it revealed that no single grouping of species was supported at the conventional (95 per cent) significance level. The contrasts analysis was therefore applied to each of a set of trees obtained by bootstrapping. The combined results indicate that the three separate contrast correlations are all significant.

The low robustness of the *Phylloscopus* tree may prove to be rather typical.



FIG. 1 Birds in the hand — a duo of leaf warblers. (Picture by Trevor Price.)

One message of the paper is that a low degree of phylogenetic certainty does not prohibit a successful comparative study. This is an area that needs further investigation, however, particularly as the bootstrapped trees are not themselves altogether independent, so compromising the tests of significance for the

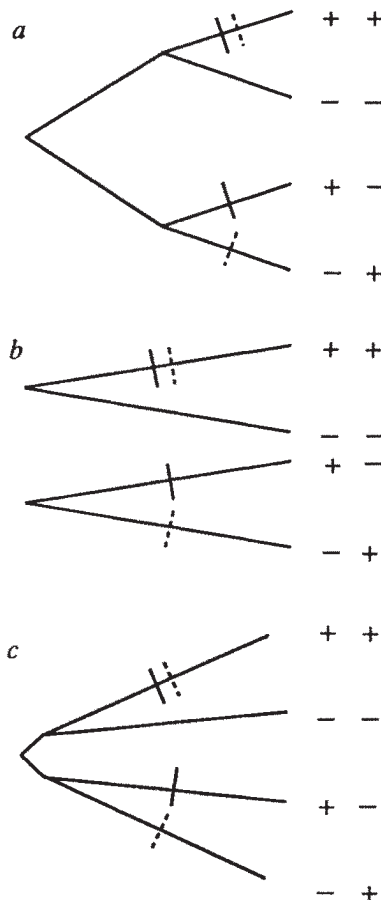


FIG. 2 Alternative patterns of relationship between two variables as might be found in phylogenetic analyses: a, typical tree; b, paired-species analysis; c, relatively simultaneous radiation.

mean contrast correlations. It would also be reassuring to see what the distribution of contrast values would be for completely arbitrary trees.

A potential problem in the use of a phylogeny from any single DNA sequence, and mitochondrial DNA in particular, is that it need not necessarily reflect the pattern of speciation itself¹⁰. But errors arising this way would presumably tend to reduce the strength of true contrast correlations rather than to create them as artefacts, so in this respect the approach is statistically conservative.

It is sometimes possible to control for ancestry without having to use an explicit phylogeny by using a paired-species analysis¹¹. In this, pairs of related species are identified which differ with respect to a variable of interest (b in Fig. 2). If enough such pairs can be iden-

tified then probabilities of association can be obtained assuming phylogenetic independence. In this way it has been shown that, as predicted, birds breeding at higher density have significantly higher rates of both intra-pair and extra-pair copulations than more solitary species¹².

One criticism of the comparative approach is that attempts to increase statistical independence may reduce the sample size and make it harder to reject an incorrect hypothesis. There is then a danger that fervent statistical rigour might obscure interesting patterns; after all, Darwin and his successors managed very well without worrying about confounding variables (but see ref. 13). Broad and even nonstatistical analyses therefore remain a good starting point for looking at evolutionary trends, although ultimately they are not totally convincing. It is noticeable, however, that phylogenies such as that for the warblers often point to ancient, quick bursts of speciation. Although this might result artefactually from a loss of resolution due to the number of nucleotide substitutions approaching saturation, it could be that it is a real response to irregular and major biogeographical events. This could mean that divergence is often really rather umbrella-like (c in Fig. 2): if the traits of interest evolve randomly through time, then within-group analyses might not in any case be so confounded.

So which method should be used? The paired-species procedure certainly has considerable advantages, especially for studying features that have evolved in many different lineages. But where interesting traits feature in only a restricted range of taxa, or when adaptive radiation is itself being investigated, it is probably time to have a rest from worrying about statistics and to invest in a lab coat instead. □

Terry Burke is in the Department of Zoology, University of Leicester, Leicester LE1 7RH, UK.

1. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford University Press, 1991).
2. Brooks, D. R. & McLennan, D. A. *Phylogeny, Ecology, and Behavior* (Chicago University Press, 1991).
3. Richman, A. D. & Price, T. *Nature* **355**, 817–821 (1992).
4. Felsenstein, J. *Am. Nat.* **125**, 1–15 (1985).
5. Ridley, M. *The Explanation of Organic Diversity: The Comparative Method and Adaptations for Mating* (Oxford University Press, 1983).
6. Höglund, J. *Am. Nat.* **134**, 72–87 (1989).
7. Kocher, T. D. et al. *Proc. natn. Acad. Sci. U.S.A.* **86**, 6196–6200 (1989).
8. Taberlet, P. et al. *Plant molec. Biol.* **17**, 1105–1109 (1991).
9. Sessions, S. K. & Larson, A. *Evolution* **41**, 1239–1251 (1987).
10. Pamilo, P. & Nei, M. *Molec. Biol. Evol.* **5**, 568–583 (1988).
11. Birkhead, T. R., Atkin, L. & Möller, A. P. *Behaviour* **101**, 101–138 (1987).
12. Birkhead, T. R. & Möller, A. P. *Biol. J. Linn. Soc.* (in the press).
13. Ridley, M. *Trends Ecol. Evol.* **7**, 37 (1992).

Spread the word

GRAFFITI, the revenge of the powerless against the powerful, disfigure ever more of our public space and transport. These scrawls are hard to remove, and no surface is safe from them. In the absence of any social solution, Daedalus has a technical fix, inspired by borophosphate-polymer glass. This material resists surface contamination. The insides of water pipes are often coated with it, as are anti-fouling surfaces at sea. Its surface partially hydrolyses to oligomer fragments which, by breaking and reforming chemical bonds to the material beneath, slowly move around. Nothing can adhere for long to such a shifty, fluid surface.

DREADCO chemists are extending this principle. They are devising other polymers whose surfaces hydrolyse or depolymerize to mobile molecular fragments. Their goal is a sort of liquid-crystal surface: hard and firm towards indentation, but viscously fluid to sideways motion. Such a surface will be always on the move, driven by air-currents, accidental contacts, or simple brownian diffusion. A sprayed-on drawing or comment will be stirred as well. Almost before the artist has completed his creation, it will begin to blur, distort and spread. In a short while it will be evenly distributed over the whole surface, as a mild and innocent colour cast.

DREADCO's 'Shiftycoat' will revolutionize public architecture. Sprayed or daubed slogans will rapidly blur away, disintegrating in great swirls and whirls as the slowly moving surface convects and dilutes them into uniformity. Some graffiti artists will be discouraged; others will welcome this endless renewal of their creative canvas; but all will be transformed from anti-social nuisances into useful citizens. They will be painting and renovating the public facilities, for free.

The colour of this redecoration service will change slightly with each added slogan, but will probably average to a rather chic grey. The marks and scuffs left by hurrying humanity will also be uniformly spread, replacing the general tattiness of public places with a permanent cheery smartness.

Householders and interior decorators will welcome Shiftycoat too. No longer will the clean patch behind a recently moved picture betray the dirtiness of the whole room. No longer will convective dust marks slowly intensify about air ducts and radiators. The dirt will be evenly distributed, giving the place that 'just-decorated' look. Even a can-of paint thrown at the wall in a domestic squabble would be slowly spread all over it, to fade and be forgotten with the argument that propelled it.

David Jones

No effect of RDGS peptides

SIR — Peptides containing the sequence Arg-Gly-Asp-Ser (RDGS), which, in vertebrates, inhibit the binding of fibronectin and other ligands to their receptors¹, have been reported by Naidet *et al.*² to interfere with gastrulation or the establishment of the dorsoventral axis in *Drosophila*. But the known *Drosophila* integrins are not required for gastrulation or dorsoventral axis formation³. In view of the facts that one of the genes responsible for organizing the dorsoventral axis in the *Drosophila* embryo encodes a serine protease with the sequence RGDS in its active site⁴, and that dorsoventral axis formation can be inhibited by the injection of protease inhibitors⁵, we attempted to repeat the experiments in which RGDS-peptides are injected into *Drosophila* embryos to test whether they act in the same way as protease inhibitors.

We found that the peptide GRGDSP had no effect on *Drosophila* gastrulation or dorsoventral axis formation, even when we used peptide solutions with concentrations up to 100 mg ml⁻¹. These solutions were simultaneously tested and shown to be active (at 1 mg ml⁻¹) in a *Xenopus* cell culture system where they inhibited the adhesion and spreading of mesodermal cells on fibronectin. The peptide was injected at different times and into different parts of the embryo, testing in each case at least 200 embryos. Most injected embryos developed normally through gastrulation.

We tested various aspects of these embryos' development. They formed a normal ventral furrow and extended their germ band normally during gastrulation. Neither of these processes occurs in genetically or experimentally dorsalized embryos. We fixed injected embryos at early gastrulation stages and stained them with antibodies against the twist protein as an assay for changes of the dorsoventral fate map. These embryos were indistinguishable from wild-type embryos. Finally, we prepared cuticles from injected embryos that had been left to complete embryogenesis. We found no signs of defective gastrulation or of dorsalization in these cuticles. Neither when protease inhibitors are effective (nuclear division cycle 6–10), nor at the time RGDS peptides were reported to be active (cycle 12–13) did the peptide affect development. It also made no difference whether the peptide

was injected into the periplasm (the cytoplasm near the egg surface) or the perivitelline space (the liquid-filled space surrounding the egg, in which the protease is thought to act, and in which ligands for cell surface receptors would be expected to be found).

Thus the dorsoventral system and gastrulation are not inhibitable by RGDS. Indeed, as we were unable to observe

any effects of RGDS peptides (with the exception of minor distortions at the anterior and posterior poles of the embryo), the effects observed by Naidet *et al.* might have to be ascribed to something other than RGDS peptides.

MARIA LEPTIN

BARBARA GRUNEWALD

DAVID STEIN

Max-Planck-Institut für
Entwicklungsbiologie,
Spemannstr. 35,
7400 Tübingen, Germany

Protein engineering in haemoglobin

SIR — The paper by Komiyama *et al.*¹ on the structural and functional significance of the D helix in human adult haemoglobin (HbA) has important evolutionary implications, particularly concerning assembly of proteins into tetrameric structures.

In HbA, O₂ binding at the haems is modulated by functional interactions with heterotropic ligands such as protons, inorganic and organic ions. Temperature also has a large effect on the O₂ affinity of HbA so that heat absorption and release can be considered a physiologically relevant modulating factor, similar to heterotropic ligands. Many interesting adaptation mechanisms involving the interplay of temperature with heterotropic ligands may be missed when experiments are performed under only one set of conditions and a single temperature.

Human fetal haemoglobin (HbF) displays at 20 °C a lower affinity for O₂ than HbA in the absence of organic phosphates². The physiologically important reverse situation is achieved at 37 °C upon addition of 2,3-diphosphoglycerate (DPG), whose effect on HbF is related to some amino-acid substitutions present in γ chains³. However, the difference in O₂ affinity observed at 37 °C is not solely due to the different modulation power of DPG with respect to HbA and HbF. In fact, when different experimental conditions are taken into account, new aspects linked to the interplay of temperature and organic phosphates are revealed. The effect of DPG on HbF renders almost identical the O₂ affinity of the haemoglobins at 20 °C, abolishing the difference observed in the absence of the effector. On going from 20 to 37 °C, by virtue of the lower overall heat of oxygenation (ΔH) displayed by HbF when in the presence of DGP ($\Delta H = -5.5$ kcal mol⁻¹ of O₂ for HbF and -8.7 kcal mol⁻¹ HbA at pH 7.4 and corrected for the heat contribution of O₂ in solution), HbA shows a lower O₂ affinity, as should be the case if O₂ has to be transferred from maternal to fetal blood.

We would like to comment on the

experimental data reported by Komiyama *et al.* in the light of this striking example. These authors make the important contribution that the presence or the absence of D helix does not affect assembly at α and β chains into cooperative tetramers, implying that this assembly and the loss of the D helix in α subunits could have been independent events in haemoglobin evolution. But as far as the functional properties are concerned, the experiments reported by Komiyama *et al.* show only that the haemoglobin molecule does not need the D helix on the β subunits under the conditions used (which unfortunately are not indicated). Hence we still do not know if the absence of the D helix induces some important alterations. At this stage it is too early to say that "the role of the D helix may simply be to link the C and E helices" or that "it is remarkable that the loss or addition of the D helix has such a small effect on O₂ binding properties and stability". A detailed functional characterization taking into account the whole network of functional interactions is needed.

B. GIARDINA, R. SCATENA

M. E. CLEMENTI, L. CERRONI,

S. G. CONDÒ

Department of Experimental Medicine
and Biochemical Sciences,
University of Rome "Tor Vergata",
00173 Rome, Italy

M. NUUTINEN

Department of Pediatrics,
University of Oulu, Oulu, Finland

O. BRIK

Zoological Laboratory,
University of Bergen, Bergen, Norway

NAGAI *ET AL.* REPLY — Giardina *et al.* discuss an interesting instance of the effect of temperature on the oxygen affinity of two haemoglobins, HbA and HbF. Because HbF has a lower oxygen affinity than HbA at 20 °C, but a higher oxygen affinity at 37 °C in the presence of 2, 3 diphosphoglycerate, Giardina *et al.* suggest that our conclusions¹ regarding the functional role of the D helix from experiments carried out at only one temperature were premature. We would

1. Hynes, R. *Cell* **48**, 549–554 (1987).

2. Naidet, C., Semeriva, M., Yamada, K. M. & Thiery, J. P. *Nature* **325**, 348–350 (1987).

3. Leptin, M., Bogaert, T., Lehmann, R. & Wilcox, M. *Cell* **56**, 401–8 (1989).

4. Chasen, R. & Anderson, K. V. *Cell* **56**, 391–400 (1989).

5. Stein, D. & Nusslein-Volhard, C. *Cell* **68**, 429–440 (1992).

like to make two points.

(1) The issue we addressed in this experiment was how the transition from non-cooperative monomeric globins to cooperative tetrameric haemoglobin may have taken place in the course of vertebrate haemoglobin evolution. Neither α nor β globin shows cooperative oxygen binding even though these proteins are assembled into dimers or tetramers in solution. Cooperativity arises only from assembly into tetramers consisting of two α and two β chains. In haemoglobin evolution, did the α and β chains diverge after globin had been assembled into cooperative homo-multimers, or did cooperativity arise from the assembly of two globins that had already diverged? The absence of D helix is the most distinct conserved feature which distinguishes modern α and β globins and the principal conclusion of our paper¹ was that cooperative tetramer can be assembled with or without the D helix. The experiment carried out at one temperature was enough to demonstrate the important conclusion that the deletion of D helix is likely to have been an independent event from assembly into cooperative tetramers.

(2) The conditions used for measuring oxygen equilibrium curves, which were unfortunately omitted from our original paper, were 50 mM bis-Tris buffer, pH 7.4, 0.1 M Cl⁻, 25 °C. These conditions closely match those found within the red cell and are widely used as standard conditions for this type of experiment. In the case discussed by Giardina *et al.*, the variation with temperature of oxygen affinity is considerably increased by the addition of DPG. This is not surprising as HbF has changes such as His 143 to Ser at the DPG binding site that weaken DPG binding. The well-characterized DPG-binding site is far removed from the D helix, so it is unlikely that the DPG effect is altered by the D helix. Cooperativity is a sensitive indicator of changes to the allosteric balance between the low-affinity T state and the high-affinity R state of the protein. Our results show that functionally important C and E helices can be connected either by a loop or a small helix without disturbing this fine balance.

KIYOSHI NAGAI
NOBORU KOMIYAMA

MRC Laboratory of Molecular Biology,
Hills Road,
Cambridge CB2 2QH, UK

JEREMY TAME

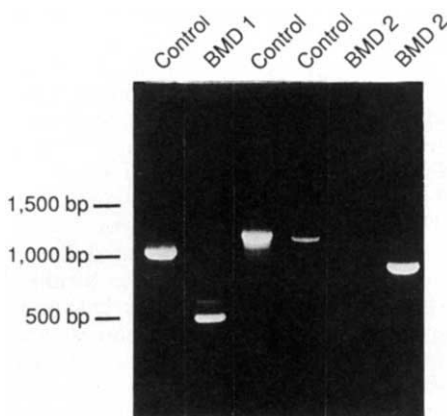
Department of Chemistry,
University of York,
Heslington, York YO1 5DD, UK

1. Komiyama, N. H. *et al.* *Nature* **352**, 349–351 (1991).
2. Doyle, M. L. *et al.* *Biochem. J.* **260**, 617–619 (1989).
3. Tomita, S. *J. Biol. Chem.* **256**, 9495–9500 (1981).
4. Perutz, M. F. *Mechanism of Cooperativity and Allosteric Regulation in Proteins* (Cambridge Univ. Press, 1990).

Dystrophin mRNA in lyophilized tissue

SIR — Khurana *et al.*¹ have recently shown that biopsy tissue can be preserved through lyophilization for later antibody-based diagnostic testing for Duchenne or Becker muscular dystrophy (DMD/BMD). We have found that lyophilized specimens can also be stored at room temperature for extended periods of time for eventual RNA-based diagnostic testing for this disease.

Cryostat sections (20- μ m) were cut from three biopsies to yield a total of approximately 25 mg of each tissue. The sections were flash-frozen in liquid nitrogen and lyophilized overnight under constant vacuum as described in ref. 1. Two of the biopsies were from BMD patients with dystrophin gene deletions previously identified by DNA analysis and the



Alterations in RT-PCR products from lyophilized samples. Lane 1, normal sized product amplified from the 5' end of the dystrophin gene. Lane 2, reduced-sized product amplified from patient 1 using the same primer set, indicating the absence of exons in this region. Lanes 3 and 4, normal products from amplification with two overlapping primer sets within the rod region of the dystrophin gene. Using the same two overlapping primer sets, lanes 5 and 6 show the missing product from one primer set and the reduced sized product from the other primer set, indicating a deletion lying somewhere between exons 44 and 52 in the patient's dystrophin gene.

third was obtained from an individual with no dystrophin gene deletion. The lyophilized tissue samples were coded to allow the RNA analysis to be performed blind, with no indication of the sample identity or the deletions present in the patient samples.

Poly(A)⁺ RNA was extracted from the lyophilized samples after 7 days' storage at room temperature. The RNA was reverse transcribed to obtain complementary DNA for polymerase chain reaction (PCR) amplification. Primers used in the PCR amplification were designed as overlapping sets to amplify the entire 14-kb dystrophin sequence (M.S.A. *et al.*, manuscript in preparation; sequences available upon request).

Following two rounds of amplification with nested primers, the PCR products were separated on agarose gels and analysed for differences from the expected pattern (see figure).

One sample was found to have the normal pattern of bands, indicating no deletion in the individual's dystrophin gene. The other two samples had deletions in two different regions. One had a deletion at the amino terminus lying between exons 2 and 7, and the second had exons missing in the rod region of dystrophin lying between exons 44 and 52. Following the RNA analysis, the sample codes were revealed and the results corroborated by comparison with DNA deletion results previously obtained for these individuals. By DNA analysis, exons 3–7 were deleted in the first sample and exons 45–47 were deleted in the second sample.

The use of RNA reverse transcription and PCR analysis (RT-PCR) provides insight into the location and extent of dystrophin gene deletions in DMD and BMD patients^{2,3}. It offers a higher level of resolution than Southern or multiplex DNA analysis and can identify mutations not detectable at the DNA level.

The RT-PCR technique is also applicable in the diagnosis of other genetic diseases. The use of lyophilized tissue samples should increase the availability of this powerful technique. Lyophilization of sample material obviates the need to transport frozen samples for analysis, which is costly, difficult to do from remote areas, and sometimes results in specimen loss because of thawing. Because the lyophilized tissues can be shipped at room temperature, this preservation method gives hospitals and physicians throughout the world access to the few specialized centres offering RNA-based testing techniques.

MARYDILYS S. ANDERSON
LOUIS M. KUNKEL

Division of Genetics and
Howard Hughes Medical Institute
at Childrens Hospital,

TEJVI S. KHURANA

Program in Neuroscience,
Harvard Medical School,
300 Longwood Avenue, Boston,
Massachusetts 02115, USA

1. Khurana, T. S. *et al.* *Lancet* **338**, 448 (1991).
2. Chelly, J. *et al.* *Nature* **333**, 858–860 (1988).
3. Roberts, R. G. *et al.* *Lancet* **336**, 1523–1526 (1990).

Scientific Correspondence

Scientific Correspondence is intended to provide a forum in which readers may raise points of a scientific character. They need not arise out of anything published in *Nature*. In any case, priority will be given to letters of fewer than 500 words and five references. □

Shy but fearless genius

I. M. Roitt

The Seeds of Time: The Life of Sir Macfarlane Burnet. By Christopher Sexton. Oxford University Press: 1992. Pp.300. £19.95.

FRANK Macfarlane Burnet's ideas put him into the genius bracket as a scientist, and in particular as a theoretical immunologist. He was also a prolific author of tremendous breadth and clarity, director of the renowned Walter and Eliza Hall Institute for Medical Research in Melbourne from 1944 to 1965, and the most honoured of all Australian scientists, being elected to the Royal Society of London and the US National Academy of Sciences, and receiving the Order of Merit, the Nobel prize and three knighthoods.

From the few occasions that I met Burnet, I formed the impression that he was somewhat puritanical. It was quite by chance that I opened the book on a page describing his Presbyterian upbringing, which groomed him for a life of self-restraint, respectability and integrity, without encouraging him to break free of his timorous, shy and introspective character.

Burnet (1899–1985) graduated in medicine from the University of Melbourne, and then spent three years working as a pathologist at the Hall Institute and two year-long visits studying bacteriology in London. He was to pursue the rest of his career in Melbourne. Early on as a virologist, he made the outstanding discovery of lysogeny, in which he recognized the ability of a bacteriophage to become noninfectious through incorporation into the hereditary constitution (genome) of the infected bacterium; much later he became convinced, correctly, that recurrent herpes associated with high antibody levels was due to reactivation of a latent virus. He showed that there was more than one strain of poliomyelitis virus, a finding that has proved to be important for control of the disease, and he did some preliminary experiments in which he attempted to grow the polio virus in tissue culture. Had he been able to continue with these experiments, he might have beaten John Enders' Nobel prizewinning team by about ten years.

He also identified the agent of Q fever. A bacteriologist, noting the fever in workers in abattoirs, tracked down a possibly infectious agent in guinea-pig spleen; this was passed to Burnet, who rapidly showed, using a simple microscope, that mice inoculated with the infectious material had inclusion bodies similar to the rickettsia, which he had so brilliantly shown to be a cause of psittacosis.

The organism was eventually named *Coxiella burneti*, so immortalizing Burnet's name in taxonomic archives. I was amused to see that he was annoyed to have missed the phenomenon of influenza haemagglutination, although he did later identify the enzyme from the vibrio that destroys the influenza receptor.

Outstanding though these findings were, they did not approach the heights



Burnet — Nobel prizewinner in 1960.

reached in his exposition of the concepts of acquired immunological tolerance and the clonal selection theory. His thoughts turned to the question of how the immune system could distinguish between molecules of the host's own body and those of an invading infectious organism. Burnet bravely hypothesized that the recognition of self occurs during the embryonic development of the immune system. A negative signal is implanted that prevents subsequent immune recognition on contact with potentially antigenic molecules later in life when the cells become immunocompetent.

Thus, in 1949, in the second edition of his book with Frank Fenner entitled *The Production of Antibodies*, Burnet predicted that "if in embryonic life, expendable cells from a genetically distinct race are implanted and established, no antibody response should develop against the foreign cell antigen when the animal takes on an independent existence". At that time, Peter Medawar was trying to understand why skin grafts between dizygotic twin cattle were accepted as readily as between monozygotic twins. Burnet and Fenner's book enabled him to make sense of this finding, and with

Rupert Billingham and Leslie Brent he went on to demonstrate acquired immunological tolerance to allogeneic skin grafts in mice injected neonatally with donor cells. It was for this work that Burnet and Medawar shared the Nobel prize in 1960; they were utterly different individuals, Burnet reticent and unworldly, Medawar the supremely articulate and urbane Englishman.

But Burnet thought his greatest achievement was his theory of clonal selection as a basis for immune responses. Neils Jerne had already proposed a darwinian natural-selection theory of antibody production. Rather than antigen acting as a template around which antibodies would be newly formed, Jerne postulated an enormous variety of pre-existing antibodies of different specificities with antigen selecting the complementary antibody, which then signals the synthesis or reproduction of molecules identical to those that had combined with antigen. It was Burnet who suggested that if each cell made one antibody and this were posted as a receptor on the cell's surface, then antigen combining with that receptor would stimulate the cell and cause it to multiply to produce a clone capable of making large amounts of that one antibody. This hypothesis has long since been accepted as one of the key ideas in modern immunology.

As director of the Hall Institute, Burnet nurtured an outstanding group of scientists including Gus Nossal, Don Metcalfe, Jacques Miller, Gordon Ada, Ian Mackay and Frank Fenner. Although he had a widely known disdain for clinical science, he did establish and strongly support the clinical research unit at the institute. But he did not put sufficient effort into keeping the institute up to date with the latest technology; in 1957, perhaps aware of the increasing need for technological innovation in the study of virology, which at that time dominated the institute's research effort, he announced peremptorily that the total energies of the institute would henceforth switch to immunology. Whatever his motives, this was a very wise move.

From his lofty perch as a Nobel prizewinner, Burnet was not reticent about holding forth on any number of different topics. In some of them he had fascinating and stimulating things to say. For example, he was very much aware of the urgent need to control human population growth and of the necessity for maintaining ecological balance, not polluting the atmosphere and using renewable energy sources. But he could also engender widespread fury by his unrestrained, simplistic and naive treatment of complex philosophical social problems. For example, he had disturbing eugenic ideas, such as the compas-

sionate infanticide of children with irremediable or serious disabilities or, even worse, the killing of violent individuals because of their danger to society. He also disparaged molecular biology, the nature and potential of which he did not really understand. And he suffered from the common affliction of eminent scientists late in their careers in believing that science had now come to an end.

Sexton, a barrister-at-law, has painted a convincing portrait of Burnet and is to be congratulated on his generally accurate exposition of the scientific issues. Personally, I would like to have seen a greater emphasis on the descriptions and

implications of self-tolerance and clonal selection at the expense of detailing Burnet's more trivial encounters with royalty. Nevertheless, the book brings alive the nature of the scientific quest and shows the excitement that a man of Burnet's genius can bring to the evolution of broad ideas, as well as how these ideas can spark off whole generations of scientists. Would that we could all have such an impact. □

I. M. Roitt is in the Department of Immunology and Rheumatology Research, University College and Middlesex School of Medicine, Arthur Stanley House, 40–50 Tottenham Street, London W1P 9PG, UK.

Electric fishnets

Max Westby

Neural Nets in Electric Fish. By Walter F. Heiligenberg. MIT Press: 1991. Pp. 179. \$50.50, £33.75.

FIRST a word of warning that may come as a disappointment to the artificial-intelligence community and a relief to others. This book has *nothing* to do with neural nets. No hidden layers or back-propagation here, this is strictly neural circuitry. The title is misleading given current usage and has already led to the inappropriate inclusion of this book in a series of strictly computational neuroscience publications.

It is 15 years since the publication of Heiligenberg's last monograph on electric fish, and a great deal of research has subsequently been carried out on these fascinating animals. His latest book is even more narrowly focused on the highly esoteric piece of behaviour known as jamming avoidance, and deals nearly exclusively with one genus, *Eigenmannia*.

Several species of fish use their electrical sensitivity not only to detect distortions of self-generated weak electric fields for object detection and navigation, but also to detect other electric fish in electrocommunication. These two functions can come into conflict — the jamming avoidance response has evolved to overcome this. The response was first described in high-frequency-wave-type species by A. Watanabe and K. Takeda in 1963. They showed that individual fish will shift the frequency of their normally invariant quasi-sinusoidal, roughly 500-hertz discharge away from that of another fish (or a signal generator) discharging within a few hertz of their own frequency. They hypothesized that this

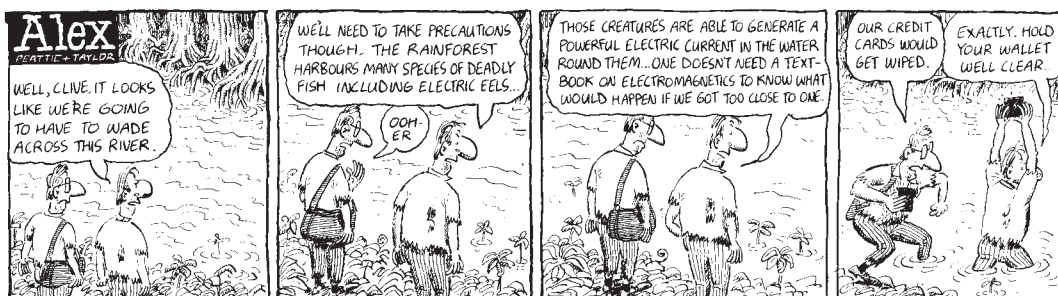
behaviour enabled the fish to avoid interference by maintaining a private channel for electrolocation.

Ten years later, Heiligenberg came on the scene with an engineering approach in which he used Fourier analysis to measure the gain and phase of electrolocating fish tracking swinging objects in the presence of jamming electric fields. Only then did it become possible to move beyond the supposition that jamming stimuli caused a deterioration in tracking performance and to characterize exactly the nature of the response. This

comparisons) at the two ears, although our acuity is an order of magnitude worse than that of *Eigenmannia*.

Perhaps the classic case of a widespread principle in neuroscience for which electric fish are an extremely good model is the principle of 'efference copy'. This principle is used extensively in the mammalian central nervous system, particularly in motor control. Here, a comparison of the intended behaviour (motor command) with actual results, through sensory feedback, is a crucial component of sensorimotor learning such as skilled reaching, grasping and locomotion.

Similarly, African mormyrid electric fish use a copy of their efferent discharge command signal to elaborately gate the electrosensory input. In this way, the fish can sort out its own self-induced electrical input from that generated by its neighbours. Unfortunately, this phenomenon is not covered in Heiligenberg's book because *Eigenmannia* does not use it. Indeed, it is the fact that this fish does not use efference copy that makes the study of its jamming avoidance response so challenging. To avoid jamming, the fish must be capable of determining both the magnitude and the sign of the difference between the



Charge card: Yuppie Alex's conclusions concerning the neuroethology of the electric eel.

rigorous approach, coupled with elegant and innovative experimental techniques, is the hallmark of his work. As it developed, unsuspecting biologists soon had no choice but to grapple with such concepts as trajectories in the amplitude-phase plane.

In this book, Heiligenberg basically tells the history of his prolific research output over the past 20 years. He is at pains to justify the narrow focus of the book, but succeeds in doing so. The phase-comparison system of *Eigenmannia* can resolve time differences of one microsecond, a remarkable feat for any nervous system, given the amount of jitter inherent in neural transmission, and a problem of general importance for other sensory systems, especially hearing, where high temporal acuity is vital for the recognition of species-specific sounds, including the distinctive features of human speech. The spatial localization of sound sources also requires time-difference measurements (among other

frequency of its own and its neighbour's discharge. The goal of the research is thus to understand the neuronal mechanism behind the response. The second half of the book is devoted to this topic and is riveting reading.

Heiligenberg takes one through this research as though it were a detective story, introducing one with enthusiasm to the battery of behavioural, electrophysiological and anatomical techniques that have allowed the almost complete description of the neuronal basis of a species-typical behaviour. This experimental work will be remembered as a classic of modern neuroethology along with that on the discrimination by toads between predators and prey, and the book is an elegant example of the power of the neuroethological approach at its very best. □

Max Westby is in the Department of Psychology, University of Sheffield, Sheffield S10 2TN, UK.

All the world as our stage

Cynthia Rosenzweig and
Daniel Hillel

The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Past 300 Years.

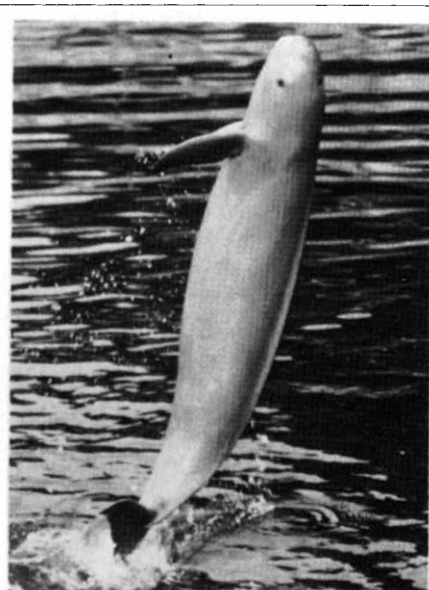
Edited by B. L. Turner II, W. C. Clark, R. W. Kates, J. F. Richards, J. T. Mathews and W. B. Meyer. *Cambridge University Press with Clark University*: 1991. Pp. 713. £75, \$100.

A HERCULEAN effort by a large group of geographers, centred at Clark University, has culminated in this massive book of 42 information-packed chapters. With some 85 authors, and a foreword, preface, postscript, as well as editorial introductions to each of the four sections, this is indeed an *opus magnum*. The selection of topics is comprehensive, the authors are leading authorities and there is an extensive bibliography. Altogether, the book is an impressive and highly laudable undertaking by geographers to 'put it all together', that is, to establish a theoretical framework for assessing the environmental changes wrought by modern societies in the past three centuries, and to document these changes in detail. As such, the tome will serve as an indispensable reference and resource on global change. It will be used by all scholars, teachers and students concerned with the increasingly pressing issue of the environment and its management — or mismanagement — by humans everywhere on our planet.

The volume is actually an expanded 'proceedings' of a large conference held in 1987 at Clark University in Worcester, Massachusetts. It is a revision of the earlier examination of the subject by geographers in 1955 that resulted in the book *Man's Role in Changing the Face of the Earth*, which was in turn inspired by G. P. Marsh's pioneering and classic work, *The Earth as Modified by Human Action*, first published more than a century ago (1885).

Geographers see it as their role to collect and integrate fragmentary data from various fields, and to reconcile the disparate points of view offered by scientists, sociologists and economists. This book is therefore important for the insights it can provide across the traditional disciplinary divides. It seeks to elucidate *what* we have done and are doing to nature, as well as *how* and *why* it is done.

The integrative purpose of the volume is clearly explained in the first chapter by R. W. Kates, B. L. Turner II and W. C. Clark. Human-induced change now en-



KOALA lemur (*Megaladapis edwardsi*), now extinct, but known from fossils in Madagascar, and a finless porpoise (*Neophocaena phocaenoides*), the smallest living cetacean. These pictures are taken from the fifth edition of *Walker's Mammals of the World* (senior author Ronald M. Nowak). On its original publication in 1964, *Nature's* reviewer called the book "one of the major contributions of the century to the study of mammalian biology". It has since established itself as the pre-eminent reference work on mammals. Completely revised and updated, the latest edition now includes descriptions and illustrations of more than 1,100 genera, covering 4,000 different species. It also now contains accounts of all mammals known in historical time, that is, since about 3,000 BC. Published by Johns Hopkins University Press. Two volumes, \$64.50, £56.

compasses nearly half of the Earth's continental surface. For most components of the biosphere, more than 50 per cent of the change has occurred in the twentieth century. Proper emphasis in the volume is placed on population growth as the main driving force of global transformation, followed by technology and sociocultural behaviour. Even with the decline in the rate of population growth and its anticipated stabilization, the rapid march of technology is still likely to alter the environment in far-reaching ways. Although an overall theory of the relationship between humans and the environment has not yet been developed, the authors posit that we do now have some useful analytical tools with which to probe this relationship.

The main body of the book is a series of expositions of specific aspects of transformation, how these aspects affect different geographic regions and the ways in which they interact. Of the four main sections, the first and last ("Changes in Population and Society" and "Understanding Transformation", respectively) deal explicitly with the human dimensions of the change. The second section ("Transformation of the Global Environment") documents some of the physical and biological changes, and the third section ("Regional Studies of Transformation") provides a selection of illustrative case studies.

It is perhaps inevitable in a work as broad as this that depth is sometimes

sacrificed and that a few weaknesses should be apparent. For instance, the choice of the past 300 years seems arbitrary, and entails a sacrifice of historical antecedents and progressions. As with all multi-authored compendia, the book is somewhat uneven. The various expositions occasionally overlap and, more regrettably, leave some notable gaps. Among the topics omitted or insufficiently emphasized, for instance, is the process of land and water contamination by agrochemicals and by domestic and industrial wastes.

Many of the chapters provide only a qualitative description of environmental problems. As environmental scientists, we would have liked a more rigorous, process-oriented analysis, including the use of mathematical models capable of formulating complex environmental interactions and predicting their likely consequences under various alternative sets of initial conditions. Perhaps the fault lies in the all-too-human tendency of groups of scholars (in this case, geographers) to 'go it alone'. The inclusion of contributions by a few more physicists and biologists might have helped in this respect.

Some of the regional case studies presented are superb: the treatment of Amazonia is outstanding, as are the chapters on the Yellow River Plain and the Malay Peninsula. Yet several equally instructive cases are neglected, such as the Aral Sea disaster in Soviet Central Asia; the San Joaquin-Kesterson Reser-

voir quandary in California; large-scale development projects, such as the Jonglei Canal in the Sudd, the Narmada project in India and the East Anatolia project in Turkey; land clearing along the Himalayan foothills and flooding in Bangladesh; land degradation in the Sahel and the Horn of Africa; and saline seeps in Australia and North America. Other important topics are treated only cursorily, such as desertification, and the greenhouse effect and its potential impact on natural ecosystems, agriculture and so on.

Unfortunately, the index is rather incomplete and does not direct the reader efficiently to the topics covered by the chapters. One example is the omission of the term 'ogallala aquifer', and this despite its explicit mention in the excellent chapter on the Great Plains region in the United States.

A final lacuna is the absence of a forward-looking assessment of possible ways in which policy-makers might take effective action. The final section provides theoretical insights that are interesting in themselves, but are scarcely sufficient to lay the necessary concrete foundations for practical strategies towards remedial action. Where might we go from here? How can human societies achieve a sustainable relationship with the Earth and its biota? These questions are left unanswered.

All the minor weaknesses notwithstanding, the very attempt by the editors and authors to encapsulate what geographers currently know and understand about so complex a subject is brave and admirable. In setting out what is known, the book, *inter alia*, exposes what is insufficiently known. The book is thus not only instructive but also challenging, making it a praiseworthy and most timely contribution. □

Cynthia Rosenzweig is at Columbia University and the NASA/Goddard Institute for Space Studies, 2880 Broadway, New York, New York 10025, USA; Daniel Hillel is in the Department of Plant, Soil and Environmental Sciences, University of Massachusetts, 11 Stockbridge Hall, Amherst, Massachusetts 01003, USA.

■ Two other books relevant to the subject of this review were published last year. *Global Environmental Change* by A. M. Mannion is a broad undergraduate text examining how natural and cultural agents have transformed the Earth during the last three million years and how this knowledge might be used to plan for the future (Longman, £14.99 (pbk)). In *Leaving Eden: To Protect and Manage the Earth*, E. G. Nisbet discusses the chemical and physical processes that control our environment and the effects of human disruption of these systems, finishing with a review of ways in which the environment can be protected and restored (Cambridge University Press, £27.50, \$42.50 (hbk), £9.95, \$14.95 (pbk)). □

Bulldoggish persuasion

Colin Patterson

Thomas Henry Huxley: Communicating for Science. By J. Vernon Jensen. *University of Delaware Press and Associated University Presses: 1991. Pp. 253. \$38.50, £29.95.*

RHETORIC — the art of persuasion in speech or writing — is not commonly seen as a necessary element in the education of scientists. In the United Kingdom, where there are no professors of rhetoric, it is evidently assumed that the art so necessary for politicians will be either inborn or acquired casually, perhaps by speaking at the Oxford Union or other debating societies. In North America, rhetoric can be part of the curriculum, and the author of this book teaches courses in it at the University of Minnesota. My only previous experience with a professor of rhetoric was through the hero of Robert M. Pirsig's *Zen and the Art of Motorcycle Maintenance* (1974), whose lectures led into his obsessive search for that slippery substance "quality". The same search must lie behind this book, a "rhetorical biography" of T. H. Huxley.

Huxley (1825–95) is best known today as a member of Darwin's coterie and as 'Darwin's bulldog', the man whose pugnacity did so much to advance the cause of evolution and materialism in the years following publication of *On the Origin of Species*. But with his lectures and essays, Huxley was also perhaps the first great publicist for science. Unlike Darwin, Huxley was not born into a moneyed dynasty and much of his early career in biology was spent seeking employment that would pay enough for him and his prospective wife to live on. His method was to make himself conspicuous, by publishing as much as he could and by attending and speaking at scientific meetings; the method is surely familiar to thousands of present-day graduates in biology. But by the time Huxley found a job, in 1854, he was already a fellow and council member of the Royal Society of London, and had won a Royal Medal; these experiences are unfamiliar to more than a handful of present-day graduates in biology, and certainly to none in need of employment. Thus Huxley was in some ways part of our world, but in others part of a different one. What can we learn from this book that may help to reconcile the two, and teach today's aspirants something of quality?

The book is in no sense a connected biography, but a collection of essays linked by graceful bridging passages and

closed by a polished summary and conclusion. The reason for this arrangement seems to be that Jensen has previously published the substance of most of the individual chapters as papers in journals. Three of the eight chapters are on "rhetorical milestones" in Huxley's career (his maiden public lecture, a Friday-evening discourse at the Royal Institution in 1852; his encounter with Bishop Wilberforce at the British Association meeting in 1860; and his lecture tour of the United States in 1876), and two chapters are on "private interpersonal communication in the life of the public rhetorician" — his three female confidantes early in his career, and the X Club, a group of nine like-minded scientists including Huxley, J. D. Hooker, Herbert Spencer, John Tyndall and John Lubbock, who dined together monthly from 1864 until age and mortality extinguished the club in the 1890s. That group had a remarkable influence on late Victorian science, providing three presidents of the Royal Society, five presidents of the British Association, and playing a large part in the foundation of *Nature* in 1869. But that is history of science. The problem with discussing Huxley's rhetoric on occasions such as his first public address and the 1860 Oxford debate is that no one now knows just what was said. The book seems to me to come to life only after the US tour, of which there are ample records, and especially in the discussion of Huxley's style in his anticlerical sallies, late in life. Jensen argues persuasively that in those debates "Huxley was adopting a rhetorical stance more similar to theologians than he would probably care to acknowledge. Rhetorically he was a secular theologian."

One learns from the book that there are four ways of delivering a talk (impromptu, memorized, speaking from manuscript, extemporaneous; Huxley became a master of the fourth), and that there is no substitute for careful preparation. On writing, Huxley recommended Sheridan's remark: "Easy reading is damned hard writing." Hard work, many drafts and a well-stocked mind will all help, but in the end the style is the man: "Nothing comes of imitation but that every man's style should express his own individuality" was Huxley's recipe. In concentrating on the rhetorical aspect, Jensen says little of the scientific content of Huxley's lectures and writing. To learn, for example, why Huxley chose the topic "animal individuality" for his first public lecture, and how that talk initiated his lifelong feud with Richard Owen, one must look elsewhere. □

Colin Patterson is at the British Museum (Natural History), Cromwell Road, London SW7 5BD, UK.

Origin and evolution of the genus *Homo*

Bernard Wood

It is remarkable that the taxonomy and phylogenetic relationships of the earliest known representatives of our own genus, *Homo*, remain obscure. Advances in techniques for absolute dating and reassessments of the fossils themselves have rendered untenable a simple unilineal model of human evolution, in which *Homo habilis* succeeded the australopithecines and then evolved via *H. erectus* into *H. sapiens*—but no clear alternative consensus has yet emerged.

TRADITIONALLY, *Homo* has been associated with brain enlargement, the acquisition of culture, a reduction in emphasis on mastication as a means of food preparation and breakdown, and a bipedal gait. The australopithecines, on the other hand, are judged not to have advanced much beyond extant apes in relative brain size, to have been dependent on large postcanine teeth for processing food and to have mixed climbing with bipedalism. Any proposal claiming to have identified the earliest evidence for *Homo* needs to demonstrate a shift from the australopithecine to the hominine grade, but the closer to the cusp between the grades the greater the difficulty in distinguishing them. This review draws upon recent developments in taxonomy and systematics and focuses on the evidence for, and the controversies about, the origin and subsequent evolution of *Homo*.

H. habilis: the case for an early *Homo* taxon

A little over 25 years ago, Leakey, Tobias and Napier¹ claimed to have identified a new, and probably the earliest, species of our own genus, *Homo*. They proposed that the new taxon, *Homo habilis* from Olduvai Gorge, Tanzania, was distinct from contemporary australopithecines but acknowledged that with respect to brain and tooth size it was significantly more primitive than *H. erectus*, hitherto the oldest recognized *Homo* species. The proposal attracted strong but conflicting criticism. Some thought that *H. habilis* showed too few advanced features to separate it from *Australopithecus*^{2,3}, but others complained that some of the fossils were indistinguishable from *H. erectus*^{4,5}. Although both Leakey and Tobias subsequently modified their detailed perceptions of *H. habilis*^{6,7}, Tobias still supports its

morphological integrity and distinctiveness⁸. Additional fossil evidence^{9–13} (Fig. 1) has confirmed the existence of a fossil hominid distinct from both *Australopithecus* and *Homo erectus*^{14,15}, and yet this acceptance¹⁶ persists in the absence of a consensus about the nature and relationships of this species. Many are content for the hypodigm (the list of fossils allocated to the species) to comprise most or all the material listed in Table 1 (refs 17, 18); others opt for more restricted membership^{19–23}. Doubts have also been expressed about the very large range of morphological variation in the postcranial material attributed to *H. habilis*^{24–27}. Does the hypodigm subsume more than one species? If it is judged to be heterogeneous, which taxa are represented? Do they belong to *Homo*, or is one, or more, an australopithecine? What are the implications for the debate about the lack of 'taxonomic space' between *Australopithecus* and *H. erectus*^{4,28}? How have attempts to define the genus *Homo* fared when matched with more recent evidence about *H. habilis*? Does the definition of *Homo* need modification, and if so, how? Finally, have phylogenetic analyses clarified the relationships between *H. habilis* and other early hominid taxa?

Fossil evidence

The initial description of *H. habilis* referred to seven cranial and postcranial specimens (Box 1). Of the Olduvai fossils subsequently added to the hypodigm (Table 1), the cranium OH 24 (OH, Olduvai hominid) and the partial skeleton OH 62 are the most important. Although OH 24 was found on the surface it was assigned to Lower Bed I, beneath Tuff 1B (ref. 29). Its early

BOX 1 First evidence of *Homo habilis*

In addition to the type specimen, six specimens were assigned to *H. habilis* in the original description¹.

Type. The juvenile skeleton (OH 7) was found in 1960 at site FLKNN in Bed I. It comprises parts of both parietals, much of the alveolar process and dentition, but little else, of a mandible and at least 13 hand bones.

Paratypes. These are remains resembling the type specimen and include skull fragments and teeth (OH 4 and 6), part of an adult foot (OH 8) and an incomplete skull of an adolescent (OH 13).

Referred specimens. A collection of juvenile cranial pieces (OH 14) and the fragmented cranial vault and dentition (OH 16) of a young adult.

Howell¹⁷ proposed that OH 6, 8 and 14 be excluded from *H. habilis*; Tobias includes OH 6 and 14 in his latest list⁸, but questions the allocation of OH 8.

Features

The authors of the original report listed the following features of *H. habilis* but they are not all distinctive.

Cranial and mandibular

1. Maxilla and mandible smaller than in *Australopithecus*, but equivalent in size to *H. erectus* and *H. sapiens*.
2. Brain size greater than *Australopithecus*, but smaller than *H. erectus*.

3. Slight to strong muscular markings.
4. Parietal bone curvature in the sagittal plane varying from slight (i.e. hominine) to moderate (i.e. australopithecine).
5. Relatively open-angled external sagittal curvature to occipital.
6. Retreating chin, with a slight or absent mental trigone.

Dental

7. Incisors large with respect to those of *Australopithecus* and *H. erectus*.
8. Molar size overlaps the ranges for *Australopithecus* and *H. erectus*.
9. Canines large relative to premolars.
10. Premolars narrower than in *Australopithecus* and within the range of *H. erectus*.
11. All teeth relatively narrow buccolingually and elongated mesiodistally, especially the mandibular molars and premolars.

Postcranial

12. Clavicle resembles *H. sapiens*.
13. Hand bones have broad terminal phalanges, capitate and MCP articulations resembling *H. sapiens*, but differ in respect of the scaphoid and trapezium, attachments of the superficial flexor tendons and the robusticity and curvature of the phalanges.
14. Foot bones resemble *H. sapiens* in the stout and adducted big toe, and well-marked foot arches, but differ in the shape of the trochlea surface of the talus and the relatively robust third metatarsal.

TABLE 1 Fossil hominid remains (by site) that are either formally or informally allocated to or declared to have affinities with *H. habilis* (better-preserved specimens in bold)

Sites	Skulls and crania	Mandibles	Specimens	Teeth	Postcranial
Olduvai (OH)	6, 7, 13 , 14, 16, 24 , 52, 62	7, 13 , 37, 62		4, 6, 15, 16 , 17, 21, 27, 31, 32, 39 , 40, 41, 42, 44, 45, 46, 47, 55, 56	7, 8 , 10, 35 , 43, 48, 49, 50, 62
Koobi Fora (KNM-ER)	807, 1470 , 1478, 1590, 1805 , 1813, 3732 , 3735, 3891	819, 1482 , 1483, 1501 , 1502, 1506, 1801, 1802, 1805 , 3734	808, 809, 1462, 1480, 1508, 1814		813, 1472, 1481, 3228 , 3735
Omo	L894-1	Omo 75-14; Omo 222-2744	L26-1g; L28-30, 31; L398-573, 1699; Omo 29-43; Omo 33-740, 3282, 5496; Omo 47-47; Omo 74-18; Omo 75s-15, 16; Omo 123-5495; Omo 166-781; Omo 177-4525; Omo 195-1630; Omo K7-19; Omo SH1-17; P933-1		
West Turkana	Kangaki I site; (no specimen no. given)	—	—	—	—
Chemeron	KNM-BC 1	—	—	—	—
Sterkfontein	Stw 53 , SE 255, 1508, 1579, 1937, 2396; Sts 19	—	—	—	—

date (Table 2) thus countered criticisms that *H. habilis* was a mistaken amalgamation of earlier, more 'Australopithecus-like' and later, dentally more advanced, 'erectus-like' remains from the middle strata of Bed II (ref. 4). The partial skeleton OH 62 (ref. 10) is a potentially rich source of information about the proportions and detailed morphology of the limbs (Box 2). A fragment of a temporal bone recovered from Chemeron, a locality in the Gregory Rift Valley to the south of Lake Turkana, was not initially referred to *H. habilis*, but a substantial number of its traits were listed as compatible with it³⁰. A new assessment of this fragment cites two features of the jaw joint that are apparently unique to *Homo*³¹.

The largest contribution to the *H. habilis* hypodigm in terms of numbers (Table 1) and quality comes from Koobi Fora, another site associated with the Gregory Rift, which lies on the northeast shore of Lake Turkana; recent examinations of the homogeneity of the *H. habilis* hypodigm have focused on the crania KNM-ER 1470 and 1813 which just antedate the evidence from Olduvai (Table 2). Hominid remains recovered from the northern group of Turkana basin sediments (specifically from members G and H of the Omo Shungura Formation) have also been likened to *H. habilis*. These include a fragmented cranium¹², two mandibles and about 20 isolated teeth^{32,33} (Table 1). A cranial fragment recovered from the Nachukui Formation, on the western shores of Lake Turkana, has also been referred to *H. habilis*³⁴. The first assessment of the hominid remains recovered from Hadar in Ethiopia and Laetoli in Tanzania suggested that an early *Homo* species may be represented^{35,36}, but these specimens have now all been accommodated within a single species, *A. afarensis*³⁷.

A fragmentary skull, Stw 53, from the cave site of Sterkfontein in southern Africa, together with isolated teeth also recovered from Member 5 at Sterkfontein, have been said to resemble *H. habilis*³⁸. Several authors have proposed that hominine material^{39,40} can also be identified within Member 4 at Sterkfontein, a collection which is dominated by remains attributed to *A. africanus* (Table 1). Attribution to *H. habilis* is also among the options considered for both cranial (SK 847 and SK 27), and dental (SK 2635) remains from Member 1 at Swartkrans^{17,41,42}. Although several authors have linked SK 847 with *H. erectus*, recent studies have emphasized its affinities with Stw 53^{43,44} and thus indirectly with *H. habilis*.

Proposals that *H. habilis* has been identified at sites beyond Africa, in the Near East and Asia, have not been sustained. The hominid fragments from Ubeidiyah in Israel were listed¹ as

possible members of the *H. habilis* hypodigm, and it was tentatively suggested that remains from Indonesia assigned to *Meganthropus palaeojavanicus* may be synonymous with *H. habilis*⁴⁵; both proposals have since been abandoned⁸.

H. habilis—one species or two?

Doubts have been expressed about the taxonomic homogeneity of the Olduvai hypodigm of *H. habilis* from the outset (Table 3). The initial criticisms, that there was a temporal basis for the taxonomic heterogeneity⁴, were effectively countered by the discovery of the cranium OH 24 (see above), but Leakey still invoked time to explain at least some of the range of morphology subsumed within *H. habilis*. He cast the mandibles OH 7 and OH 13 as, respectively, the early and late representatives of a 'sapiens-like' lineage, whose morphology he contrasted with the 'protopithecantropine' features of specimens such as the cranium OH 16 (ref. 7). A subsequent assessment sorted the specimens differently, concluding that morphological differen-

TABLE 2 Best estimates of the geological age of some of the better-preserved East African fossil evidence for *H. habilis*

	Olduvai (OH)	Sites Koobi Fora (KNM-ER)	MO
Myr	1.6	13	
		16	
	1.7		
	1.8	3891	
		1805	
	24		
	1.9	1470, 1802, 1813, 3732, 3735	L894-1
	2.0		

ces between OH 13 and 24, on the one hand, and OH 16 and OH 7, on the other, hinted at taxonomic variation⁹.

The additions to the *H. habilis* hypodigm provided by the Koobi Fora remains failed to clarify the situation. Several observers^{17,18} stress the taxonomic unity of the augmented hypodigm, while others have re-emphasized the australopithecine affinities of part of the enlarged hypodigm⁴⁸⁻⁵⁰. The theme of the morphological dichotomy between OH 13 and 24, and OH 7 and 16 (ref. 9) has been taken up by others^{48,50}, who suggest that *H. habilis* should be restricted to OH 7 and 16, that is, the larger-toothed, bigger-brained, and presumably larger-bodied, component of the main hypodigm. In this scheme, specimens without these attributes, such as OH 13 and 24 and the crania KNM-ER 1805 and 1813 from Koobi Fora, were judged to be "late-surviving small *Australopithecus* individuals that were contemporaneous first with *H. habilis*, then with *H. erectus*"⁵⁰.

These latter proposals involve removing a substantial part of the *H. habilis* hypodigm from *Homo*, but others have been prepared to accommodate both subsets within the same genus. Although some support a split of the hypodigm based on brain and tooth size^{20,21,51}, among other considerations, others have suggested schemes that partition the remains geographically, with *H. habilis* confined to Olduvai, and one or more new species of *Homo* at Koobi Fora^{22,23}. Groves and Mazak designated one of these new species *H. ergaster*, but others consider this as most probably synonymous with *H. erectus*²³, or a subset thereof⁵³ (see below).

Recent research has compared variation within *H. habilis* with that within living primates in order to assess the probability that *H. habilis* samples a single species, but the results have been equivocal. A comparison of two crania from Koobi Fora (formerly East Rudolf), KNM-ER 1470 and 1813 (Fig. 2), cast doubt on a single-species solution⁵⁴; another, focusing on brain size⁵⁵, supported it. The former study concluded that the contrasts between KNM-ER 1470 and 1813 were greater in degree than equivalent differences between male and female gorilla crania, whereas the latter concluded that the endocranial volumes, 752 and 510 cm³, of the two fossil crania did not imply a degree of intraspecific variation exceeding that in sexually dimorphic higher primate species. Such considerations have become more sophisticated to the extent that separate attention has been paid to the pattern as well as to the degree of morphological variation⁵⁶. Variables are assessed as 'good' or 'poor' taxonomic discriminators, with greater emphasis placed on those variables that vary more between than within species closely related to the early hominids.

The two most recent studies of *H. habilis*^{8,53} differ in emphasis but are partly consistent in their taxonomic conclusions. In his magisterial review of Olduvai remains attributed to *H. habilis*⁸, Tobias concludes that the features originally said to characterize *H. habilis*¹ (Box 1) have stood the test of time, and has no hesitation in retaining all of the Olduvai hypodigm within *H. habilis* and including the material from Koobi Fora that has been referred to *H. habilis*; his is a single-species interpretation of *H. habilis*. My own approach was different: I submitted the cranial remains attributed to *H. habilis* to examinations designed to test the null hypothesis of a single early *Homo* species. The degree of variation in the sample was compared with that in another synchronic hominid species, *Paranthropus boisei*, as well as with the degree of variation observed in *H. erectus* and the gorilla. Patterns of variation in early *Homo* were compared with those in *H. sapiens* and *Pan troglodytes*, the two species that are most closely related to early hominids and which share patterns of cranial variation⁵⁶. In the event there were few cases in which the degree of variation in the main *H. habilis* hypodigm exceeded that in the comparative material, but there were several examples (facial shape and premolar tooth morphology, for example) in which the pattern of cranial, mandibular and dental variation differed. However, in each case the evidence for taxonomic heterogeneity came not from the Olduvai part of the hypodigm

BOX 2 OH 62—a skeleton belonging to *Homo habilis*

The partial skeleton, OH 62, was found in 1986¹⁰—more than 300 fragments were eventually recovered from the site. Fragmented skeletal material had previously been recovered from Olduvai (Box 1), but none of the earlier specimens provided such clear evidence about the overall size and proportions of the postcranial skeleton of *H. habilis*.

Skull. The palate is the best preserved part of the skull; details of the floor of the nose, the shape of the tooth row and the size of the teeth point to it belonging to *Homo* and not *Paranthropus*.

Limbs. Sufficient of the humerus and forearm bones and of the femur and leg bones are preserved to enable their lengths to be estimated and to allow the relative lengths of the limbs to be determined.

Compared with modern humans, extant African apes have relatively shorter legs and longer forelimbs. The estimated lengths and robusticity of the humerus and forearm bones of OH 62 suggest that its proportions are remarkably ape-like⁷⁰. Weight/stature relationships are also distinct between the African apes and modern humans; on these criteria, OH 62 is ape-like⁶⁸.

Body weight. Limb bone dimensions predict a body weight of at least 30 kg⁹¹.

but from remains referred to *H. habilis* from the Koobi Fora collection. This led me to agree with Tobias and accept the Olduvai hypodigm of *H. habilis* as evidence for that species, but to suggest that the Koobi Fora hypodigm sampled two species of early *Homo*, one conspecific with *H. habilis* from Olduvai and the other a new species of early *Homo*, already designated *H. rudolfensis* Alexeev, 1986 (Box 3). *H. ergaster* Groves and Mazak, 1975 is not available as the name of the second species of early *Homo* because the type specimen KNM-ER 992 resembles, though is not necessarily conspecific with, *H. erectus*. Thus I have put forward *H. ergaster*⁵³ as the proper name for the probable African precursor of *H. erectus*, a taxon which would also include the crania KNM-ER 3733 and 3883 and probably also the skeleton KNM-WT 15000, although formal assignment of that specimen must await its detailed description and analysis.

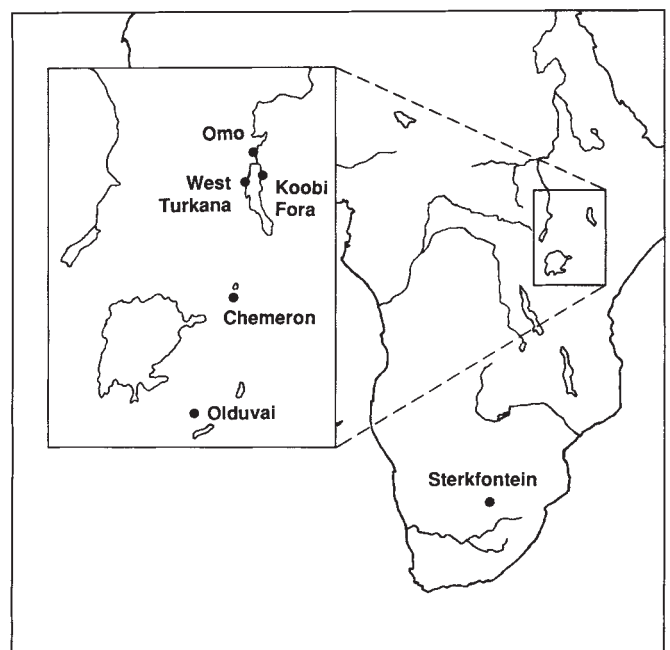


FIG. 1 Location of sites where remains attributed to, or likened to, *H. habilis* have been found.

BOX 3 Evidence for and features of *Homo habilis sensu stricto* and *Homo rudolfensis*

It has been suggested that *H. habilis sensu stricto* and *Homo rudolfensis* are species components of the larger *H. habilis* hypodigm⁵³. Specimens allocated to the two taxa are set out below.

H. habilis sensu stricto

Olduvai:

OH 4, 6, 8, 10, 13–16, 21, 24, 27, 35, 37, 39–45, 48–50, 52, 62

Koobi Fora:

KNM-ER 1478, 1501, 1502, 1805, 1813, 3735

H. rudolfensis

Koobi Fora:

KNM-ER 813, 819, 1470, 1472, 1481–1483, 1590, 1801, 1802, 3732, 3891

Skull and teeth	<i>Homo habilis s.s.</i>	<i>Homo rudolfensis</i>
Absolute brain size (cm ³)	$\bar{X}$ = 610	$\bar{X}$ = 751
Overall cranial vault morphology	Enlarged occipital contribution to the sagittal arc (but see ref. 8 for a contrary view)	Primitive condition
Endocranial morphology	Primitive sulcal pattern ⁸² (but see Holloway ⁸³)	Frontal lobe asymmetry ^{82,84}
Suture pattern	Complex	Simple
Frontal	Incipient supraorbital torus	Torus absent
Parietal	Coronal > sagittal chord	Primitive condition
Face-overall	Upper face > midface breadth	Midface > upperface breadth; markedly orthognathic

Nose	Margins sharp and everted; evident nasal sill	Less everted margins; no nasal sill
Malar surface	Vertical, or near vertical	Anteriorly inclined
Palate	Foreshortened	Large
Upper teeth	Probably two-rooted premolars	Premolars three-rooted; absolutely and relatively large anterior teeth
Mandibular fossa	Relatively deep	Shallow
Foramen magnum	Orientation variable	Anteriorly inclined
Mandibular corpus	Moderate relief on external surface	Marked relief on external surface
	Rounded base	Everted base
Lower teeth	Buccolingually-narrowed	Broad postcanine crowns
	post-canine crowns	
	Reduced talonid on P ₄	Relatively large P ₄ talonid
	M ₃ reduction	No M ₃ reduction
	Mostly single-rooted mandibular premolars	Twin, plate-like, P ₄ roots, and bifid, or even twin, plate-like P ₃ roots
Postcranium		
Limb proportions	Ape-like	?
Forelimb robusticity	Ape-like	?
Hand	Mosaic of ape-like and modern human-like features	?
Hindfoot	Retains climbing adaptations	Later <i>Homo</i> -like
Femur	Australopithecine-like	Later <i>Homo</i> -like

Evidence from limbs

Morphological characterization of the postcranial anatomy of *H. habilis*, together with its functional interpretation, have proved to be topics every bit as controversial as those involving the cranial evidence.

The first integrated account of fossil evidence of the foot and leg of *H. habilis*, like that of the cranium, was based on the evidence from Olduvai, and emphasized the ways in which these remains resembled those of *H. sapiens*⁵⁷. But functional conclusions based on more detailed descriptions of foot (OH 8, 10) and leg (OH 35) fossils, which are probably from the same individual⁵⁸, were generally more cautious. They stressed that the unique striding gait of *H. sapiens* has not yet been achieved⁵⁹ and suggested that the knee joint may have been imperfectly adapted to bipedalism⁶⁰. Despite these caveats, the inferred functional affinities between the fossils of *H. habilis* and *H. sapiens* were still emphasized, and a bipedal gait of the modern human type was widely claimed for the former species.

Subsequent reassessments of the anatomy of the OH 8 foot stressed its potential for climbing, and its retention of anatomical features seen in living non-human anthropoid apes^{58,61}. Although OH 8 apparently possessed the mechanism that transforms the foot into a rigid, close-packed, organ during the support phase of bipedal walking⁶¹, it lacks at least some of the refinements (such as lateral deviation of the heel) present in *H. sapiens*. The same author was also struck by the lack of evidence for a propulsive role for the big toe, an important factor in the bipedal gait of modern humans⁶².

A different and more complex interpretation of the hindlimb anatomy of early *Homo* has emerged from the Koobi Fora evidence, although Tobias⁸ correctly cautions that the specimens providing much of the conflicting evidence (in particular the femora KNM-ER 1472 and 1481A and the talus KNM-ER 813) may not belong to *H. habilis*, but to *H. erectus* or *H. ergaster*.

But, the features that prompted the assignment KNM-ER 1472 and 1481A to *H. erectus*⁶³ may not be diagnostic of that species⁶⁴ and are likely to be derived features of all *Homo* species, including *H. habilis*. Others have demonstrated that the two Koobi Fora specimens are distinct from australopithecine femora⁶⁵ and that a talus, KNM-ER 813, from a similar-aged horizon as the femora, resembles modern human tali much more closely than do australopithecine tali⁶⁶. Thus, there are at Koobi Fora leg fossils whose later *Homo*-like morphology contrasts with that of the more australopithecine-like morphology of the Olduvai remains. These relatively derived remains from Koobi Fora are found alongside a specimen such as KNM-ER 3735, which is judged to resemble the more primitive OH 62 skeleton⁶⁷.

An analysis of estimated stature/body weight relationships has also shown that whereas predictions based on the two Koobi Fora femora are in line with modern human and archaic *H. sapiens* relationships, they are substantially different from predictions based on the australopithecine-like Olduvai *H. habilis* remains, which instead conform to predictions based on the living African apes⁶⁸. This wide range of morphology has no apparent allometric basis, at least as far as the proximal femur is concerned⁶⁹, and is thus additional evidence for taxonomic heterogeneity in early *Homo*^{26,53}.

Evidence of the forelimb skeleton of Olduvai *H. habilis* is meagre, with the OH 7 hand providing the most pertinent information. Its fragmentary nature precludes comprehensive morphological and functional analysis, but it is generally interpreted as demonstrating a mosaic of an ape-like carpus with a thumb which could have both been rotated within the hand to allow pulp-to-pulp opposition, as well as used to provide the kind of firm support essential for effective tool manufacture and manipulation^{58,61,90}.

TABLE 3 Multiple taxon solutions for crania and mandibles attributed to *H. habilis*

Author(s)	OH	Specimens KNM-ER	Taxon names	Comments
Robinson (1965) ⁴	7	—	<i>A. aff. A. africanus</i>	C/P ₃ ratio; $\bar{C}$ morphology
	13	—	<i>H. aff. H. erectus</i>	U-shaped mandibular contour; gracile mandibular corpus
Leakey <i>et al.</i> (1971) ⁹	7, 16	—	<i>H. habilis</i>	<i>H. erectus</i> -type, evenly curved lambdoid suture
	13, 24	—	<i>H. habilis/Homo</i> sp.	<i>H. sapiens</i> -type, V-shaped lambdoid suture
Groves (1979) ⁵²	7, 13, 16, 24	—	<i>H. habilis</i>	Narrow premolars; upper face > mid face
	—	1470, 1590, 1802	<i>H. rudolfensis</i>	Mid face > upper face; P ₄ large relative to canine size
	—	730, 820, 992, 1805, 1813	<i>H. ergaster</i>	Broad premolars; P ₄ < P ₃
Leakey <i>et al.</i> (1978) ⁸⁷	7	1470, 1590, 1802	<i>H. habilis</i>	Enlarged anterior and cheek teeth; robust mandibles; cranial capacity > 750 cm ³
	13	992, 1813	<i>A. aff. A. africanus</i>	Small molars and premolars; cranial capacity approx. 600 cm ³
Stringer (1986) ²¹	7, 24	1470, 1590, 1802, 3732	<i>H. habilis</i> (group 1)	Large cranial capacity approx. 750 cm ³ ; flat lower face; large teeth and jaws
	13, 16	992, 1805, 1813	<i>H. habilis</i> (group 2) or <i>H. ergaster</i>	Small cranial capacity approx. 510 cm ³ ; projecting lower face; small teeth and jaws
Chamberlain (1989) ²³	7, 13, 16, 24, 37	—	<i>H. habilis sensu stricto</i>	Premolars and molars mesiodistally elongated; reduced jaw size relative to neurocranium
	—	992, 1470, 1483, 1802, 1805, 1813, 3734	<i>Homo</i> sp.	Mandibular premolar roots complex; shares traits with 'robust' australopithecines
Wood (1991) ⁵³	7, 13, 16, 24, 37, and so on	1478, 1501, 1805, 1813, 3735, and so on	<i>H. habilis</i>	Small neurocranium approx. 500 cm ³ ; upper face > mid face; small, buccolingually narrow P ₃ -M ₁ ; M ₃ reduction
	—	1470, 1482, 1590, 1802, 3732, and so on	<i>H. rudolfensis</i>	Large neurocranium approx. 750 cm ³ ; orthognathic with mid face > upper face; large postcanine teeth; M ₃ > M ₂

Considered with the evidence of the relatively primitive limb proportions of OH 62 (refs 10, 70), the architecture of the limbs of *H. habilis sensu stricto* is closer to that of the australopithecines than to later *Homo*.

Defining the genus *Homo*

The task would be greatly eased were agreement to be reached on definitive criteria for the genus *Homo*. Mayr has insisted that the genus is impossible to define and delimit on a purely morphological basis⁷¹, except in that *Homo* was characterized by "upright posture, with its shift to a terrestrial mode of living and the freeing of the anterior extremity for new functions which, in turn, have stimulated brain evolution". Thus defined, Mayr's genus *Homo* embraced the then known australopithecines, *H. erectus* and *H. sapiens*. Simpson⁷² subsequently warned, in the context of what he described as "the chaos of anthropological nomenclature", that genera are "necessarily more arbitrary and less precise in definition than the species", but suggested that a genus can be defined as "a group of species believed to be more closely related among themselves than to any other species placed in other genera".

The dictum adumbrated by Mayr and Simpson, that it is the species that make the genus and not *vice versa*, is borne out by critical inspection of the definitions of *Homo* provided by Le Gros Clark⁷³ and Robinson⁷⁴. The former is a concatenation of the features of what was then known of *H. sapiens*, *H. neanderthalensis* and *H. erectus*. This is reflected in some of the proposed components (such as brain size with a range of 900–2,000 cm³), but the definition also includes references to morphological trends, for example a centrally situated foramen magnum, reduced lingual cusp on the third lower premolar,

relative size reduction of the third molar, and follows Mayr to the extent that it specifies a "limb skeleton adapted for a fully erect posture and gait"⁷³. When announcing *H. habilis*, Leakey *et al.*¹ made only modest amendments to Le Gros Clark's definition of *Homo*, and Robinson later emphasized only three additional features: a nasal sill, "harmoniously proportioned" anterior and postcanine teeth and incompletely molarized crowns in deciduous first lower molars⁷⁴.

Few explicit definitions of *Homo* have been offered since, but the increasing use of cladistic methods within hominid palaeontology means that it is possible to deduce the character state changes that define the *Homo* clade. Cladistic analyses of early hominids are at present confined to cranial evidence, and surprisingly few of these analyses combine data from early hominids, later *Homo* and extant *H. sapiens*. A.T. Chamberlain and I investigated the distribution of 90 cranial, mandibular and dental characters across a range of hominid species and outgroups¹⁵ and found that the resulting *Homo* clade could be defined by 8 character state changes (Box 4). Others²⁰ have also emphasized reduced prognathism, dental reduction and brain enlargement as factors defining the *Homo* clade.

Diagnostic features of *H. habilis*

H. habilis was characterized¹ by the features listed in Box 1. Of the eleven that relate to the cranium, mandible and dentition, four—the absolute size of the maxilla and mandible, brain size, the degree of parietal sagittal curvature and molar crown size—were said to be intermediate between the condition in the australopithecines and in the then oldest hominine taxon, *H. erectus*. In two characters the hypodigm was specified as resembling existing taxa, namely occipital sagittal curvature (*H. sapiens*)



FIG. 2 Photograph of the crania KNM-ER 1470 (left) and 1813 (right). These two crania, both dating from around 1.9Myr, would need to be subsumed within a single early *Homo* species. Although they share a similarly shaped neurocranium, differences in their facial anatomy are less easy to accommodate in a single species model.

and premolar width (*H. erectus*). The reference to a retreating chin, with no menal trigone, does not suggest a significant departure from the australopithecine condition, and another character, the degree of cranial muscular marking, is too non-specific to be helpful. The diagnostic dental features are relative incisor and canine size and the shape of the tooth crowns, with the buccolingual narrowing of the mandibular premolars and molars a particularly important feature. But others dispute the uniqueness of narrow premolars among Plio-Pleistocene hominids^{4,46,75}. As for the implications of the postcranial skeleton on gait, all the remains that have been entertained as belonging to the hypodigm of *H. habilis* have been interpreted as belonging to a bipedal hominid, but the extent to which *H. habilis* was also a climber is still debated^{76,77}.

The 1964 definition was later expanded⁷⁸ and emphasized the peculiarities of the morphology of the mandibular canines and premolars of *H. habilis* and differences in the trend of mandibular premolar crown size between *H. habilis* and australopithecines. Tobias¹⁶ has cited additional "critical features of the morphology of *H. habilis*". These include the relative crown size of the second and third lower molar, vertically-thin brow ridges, minimal pneumatization of the cranium, a prognathous face, an anteriorly situated foramen magnum, a short basicranium, coronally orientated petrous bones and a "very robust" mandible with a slight chin. However, he does not indicate whether any of these features are unique to *H. habilis*. In an encyclopaedic summary of the features of *H. habilis*, he refers to 344 cranial, mandibular and dental traits, but although their expression is meticulously recorded for *A. africanus*, *A. robustus*, *A. boisei* and *H. erectus*, their wider comparative context is not explored and so their taxonomic and phylogenetic value have yet to be demonstrated.

Cladistic analyses have made little contribution to the search for distinctive features, or autapomorphies, of *H. habilis*. In an early study⁷⁹, an endocranial volume of more than 600 cm³ was cited as a character distinguishing *A. habilis* (that is, *H. habilis*) from *A. africanus*, but in a wider taxonomic context this cannot be regarded as anything other than a part of what has been called a 'combination' definition⁸⁰. Others were unable to identify any autapomorphic features⁸¹. This was implicitly conceded in another analysis⁴⁹ but the value of this latter is reduced because no distinction was made between 'gracile' australopithecines from southern Africa and the Olduvai hypodigm of *H. habilis*. A cladistic analysis based on quantitative cranial

data⁸² deduced only one *H. habilis* autapomorphy, a relatively narrow mid-face but Chamberlain and I (more recently) identified two further *H. habilis* autapomorphies using an augmented data set¹⁵—an elongated anterior basicranium and a mesiodistally elongated crown in the first upper molar. When the two

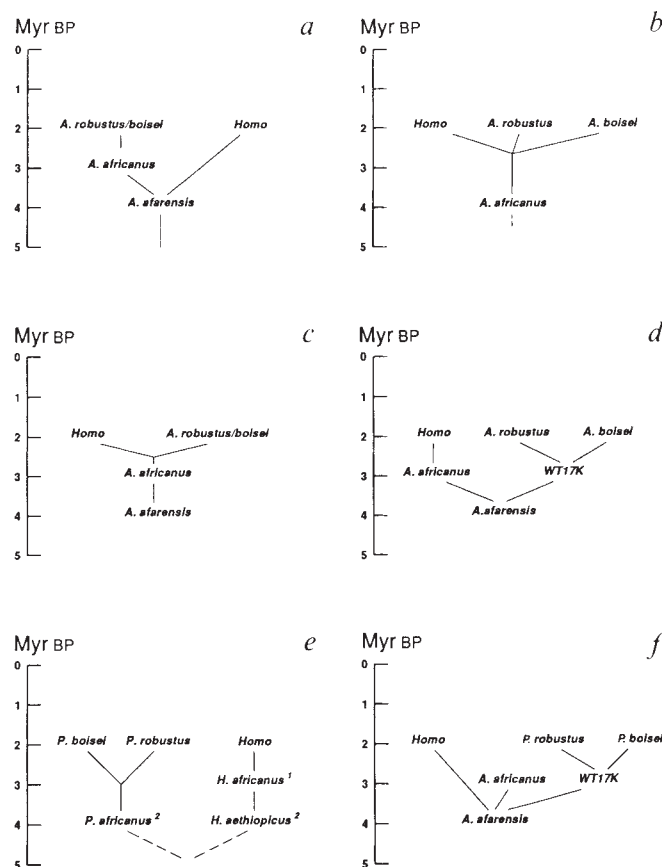
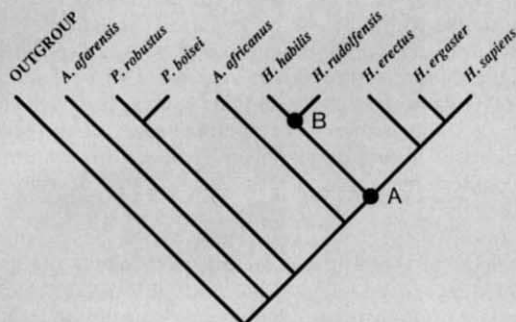


FIG. 3 Comparison of the relationships of the *Homo* lineage/clade in published early hominid phylogenetic schemes: *a*, ref. 84; *b*, ref. 88; *c*, ref. 14; *d*, ref. 89; *e*, ref. 49; (1) *A. africanus*, (2) *A. afarensis*; *f*, ref. 53.

BOX 4 *Homo* defined cladistically

CLADISTIC, or phylogenetic analysis, is a method of investigating the relationships between taxa. It makes the assumption that morphological evolution is parsimonious, that is, similar morphological features, or characters, are more likely to be derived from a shared common ancestor than to have evolved independently. Each character has a hypothetical morphocline of states, from primitive to derived. The primitive condition is assumed to be that which is most widely distributed in related taxa and/or the condition which has the least complex developmental history. Character states uniquely defining a taxon are known as autapomorphies; their distribution does not help in establishing relationships. Character states distributed within part of the cladogram (defining a 'node') are called synapomorphies, and states common to all the taxa in a cladogram

are referred to as symplesiomorphies. Each character and the distribution of its states among the taxa to be assessed allows the generation of a cladogram. The cladogram, or pattern of relationships, supported by the majority of characters is assumed to be the most probable hypothesis of relationships.

The cladogram presented above is the most parsimonious generated from a set of 90 cranial, mandibular and dental characters^{15,53}. The resulting *Homo* clade is defined by the following character state changes at node A.

1. Increased cranial vault thickness.
2. Reduced postorbital constriction.
3. Increased contribution of the occipital bone to cranial sagittal arc length.
4. Increased cranial vault height.
5. More anteriorly-situated foramen magnum.
6. Reduced lower facial prognathism.
7. Narrower tooth crowns, particularly mandibular premolars.
8. Reduction in length of the molar tooth row.

The two species comprising the original *H. habilis* hypodigm, *H. habilis sensu stricto* and *H. rudolfensis*⁵³ share a hypothetical ancestor with each other which neither shares with any other taxon. That sister group is defined by the five character state changes at node B.

1. Elongated anterior basicranium.
2. Higher cranial vault.
- 3, 4. Mesiodistally-elongated M_1 and M_2 .
5. Narrow mandibular fossa.

proposed species subdivisions of the early *Homo* hypodigm, *H. habilis sensu stricto* and *H. rudolfensis*, are included separately in a cladistic analysis, they are linked as sister taxa within a clade defined by five character states (Box 4).

Phylogenetic relationships

Phylogenetic analyses that have included *H. habilis* as a separate and unified species have almost always concluded that it is part of a clade, or monophyletic group, corresponding to the genus *Homo*, along with *H. erectus* and *H. sapiens*. In this clade, *H. habilis* is usually, but not universally, nominated as the sister taxon of *H. erectus* and *H. sapiens*^{14,15,20,46,82-84}, species which are themselves clearly defined cladistically^{13,20}. There is, however, substantially less consensus about the relationships of *H. habilis*, and the *Homo* clade, with australopithecine taxa (Fig. 3). The position of *A. africanus* is pivotal. Some have interpreted it as exclusively ancestral to either the *Homo* (Fig. 3d, e) or the 'robust' australopithecine clades (Fig. 3a), but others see it as the common ancestor of both (Fig. 3b, c). The results of two formal cladistic analyses^{14,15}, concur in that *A. africanus* is strongly associated with neither the *Homo* nor the 'robust' australopithecine clades (Fig. 3f). This confusion is caused not for the lack of features shared by *A. africanus* and either clade, but because this species represents a mosaic of relatively primitive characters and traits that are derived either

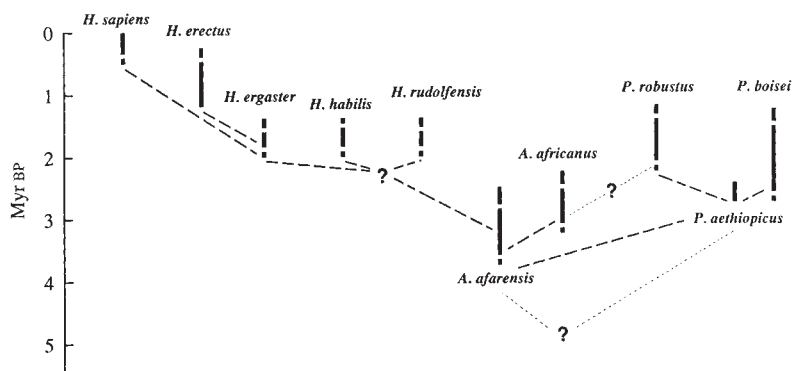
in the direction of *Homo* (such as the cranial vault), or of the 'robust' australopithecines (face and masticatory apparatus). This suggests that *A. africanus* is too specialized to be ancestral to either the *Homo* or the 'robust' australopithecine clades, or even for it to be the common ancestor of all later hominids (Fig. 3f).

Discussion

There is little doubt that the enlarged hypodigm of *H. habilis* subsumes remains that have traits more derived than those of *Australopithecus* and are yet distinct from approximately contemporary fossils usually attributed to African *H. erectus*, or by some, to *H. ergaster*. But if its taxonomic integrity is retained, it is a species that manifestly embraces an unusually large amount of variation^{21,26,51,53,54}. Those who believe this range to be unacceptably wide, and thus for whom *H. habilis* represents more than one species, disagree about how the hypodigm should be apportioned.

Most (but not all^{14,19,21}) regard the Olduvai hypodigm as taxonomically homogeneous (Table 2). The Koobi Fora evidence is either regarded as representative of a single species distinct from *H. habilis sensu stricto*⁴³, or as a taxonomically heterogeneous hypodigm that samples both *H. habilis sensu stricto* and at least one other *Homo* species⁵³, for which *H. rudolfensis* has apparent priority as a species name⁸⁵. These two early *Homo* species are sufficiently distinct for them to have substantially

FIG. 4 A phylogenetic scheme for hominid evolution based on a recent analysis of hominid taxonomy and relationships⁵³. The horizontal axis corresponds roughly to relative and absolute postcanine tooth size, so that forms with substantial tooth rows are to the right, and species demonstrating premolar and molar reduction and simplification are to the left. The phylogeny represented by the bold broken lines assumes that the two best known 'robust' australopithecine species, *P. robustus* and *P. boisei* shared a common ancestor which was not unlike *P. aethiopicus* (that is, like KNM-WT 17000). But, this monophyletic origin for *Paranthropus* is only marginally more parsimonious than deriving *P. robustus* from *A. africanus*. The similarities in facial form between *H. rudolfensis* and a probable *Paranthropus* clade are most parsimoniously interpreted as convergent features.



different grounds for their inclusion in *Homo* (Box 3). Whereas *H. habilis sensu stricto* is hominine with respect to its masticatory complex, it retains an essentially australopithecine postcranial skeleton. *H. rudolfensis*, on the other hand, probably combines a later *Homo*-like postcranial skeleton with a face and dentition adaptively analogous to those of the 'robust' australopithecines.

The earliest sound evidence for the larger *H. habilis sensu lato* hypodigm is dated to ~1.9 Myr BP (Table 3), but a series of isolated teeth from members E, F and G of the Omo Shungura Formation and the evidence of the Chemeron temporal fragment may extend its known time range to well beyond 2.0 Myr BP^{31,53,86}. The former date, 1.9 Myr BP, is little different from the age of the earliest known evidence for *H. ergaster*, the early African equivalent of *H. erectus*. If early *Homo* does subsume more than one species, the period before 2.0 Myr BP may have witnessed not merely the emergence of one hominid of the *Homo* grade, but a substantial radiation of early hominids, namely *H. habilis sensu stricto*, *H. rudolfensis* and *H. ergaster* (Fig. 4), each demonstrating a significant and distinctive shift away from the australopithecine adaptive plateau.

This multiple-species solution must now be tested. This will involve establishing sound criteria for confirming hypotheses of taxonomic heterogeneity, including refining the distinctions

between the degree and pattern of variability, as well as confirming the suitability of comparative analogues for the assessment of variation both within and between species. Other tests might also be explored, for instance, how realistic are the ranges of body weight estimated for each taxon? Do the proposed taxa imply either unacceptably large, or unreasonably small, levels of sexual dimorphism? Is there any evidence that they occupied the same, or different, parts of the palaeo-landscape?

If the multiple-species solution survives rigorous examination, other aspects of the hominid record, particularly the archaeological evidence, must be integrated with this new interpretation of the palaeontological data. Only when morphological studies, embracing both function and life history are integrated with the contextual, and particularly the behavioural, evidence will we substantially increase our knowledge and understanding of the emergence and early evolution of our own genus. A sound taxonomic interpretation of the fossil record is the foundation upon which more complex hypotheses can be constructed. □

Bernard Wood is in the Hominid Palaeontology Research Group, Department of Human Anatomy and Cell Biology, University of Liverpool, PO Box 147, Liverpool L69 3BX, UK.

1. Leakey, L. S. B., Tobias, P. V. & Napier, J. R. *Nature* **202**, 7–9 (1964).
2. Le Gros Clark, W. *Discovery* **25**, 49 (1964).
3. Holloway, R. L. *Nature* **208**, 205–206 (1965).
4. Robinson, J. T. *Nature* **205**, 121–124 (1965).
5. Brace, C. L., Mahler, P. E. & Rosen, R. B. *Yrbk. Phys. Anthropol.* **16**, 50–68 (1972).
6. Tobias, P. V. *Curr. Anthropol.* **6**, 391–411 (1965).
7. Leakey, L. S. B. *Nature* **209**, 1279–1281 (1966).
8. Tobias, P. V. *Olduvai Gorge IV: The Skulls, Endocrania and Teeth of Homo Habilis* (Cambridge Univ. Press, UK, 1991).
9. Leakey, M. D., Clarke, R. J. & Leakey, L. S. B. *Nature* **232**, 308–312 (1971).
10. Johanson, D. C. *et al.* *Nature* **327**, 205–209 (1987).
11. Leakey, R. E. F. *Nature* **242**, 447–450 (1973).
12. Boaz, N. T. & Howell, F. C. *Am. J. Phys. Anthropol.* **46**, 93–108 (1977).
13. Coppens, Y. in *Current Argument on Early Man* (ed. Konigsson, L.-K.), 207–225 (Pergamon, Oxford, 1980).
14. Skelton, R. R., McHenry, H. M. & Drawhorn, G. M. *Curr. Anthropol.* **27**, 21–43 (1986).
15. Chamberlain, A. T. & Wood, B. A. *J. hum. Evol.* **16**, 119–133 (1987).
16. Tobias, P. V. in *Hominidae: Proc. 2nd Int. Congr. Human Paleontology* (ed. Giacobini, G.) 141–149 (Jaca, Milan, 1989).
17. Howell, F. C. in *Evolution of African Mammals* (eds Maglio, V. J. & Cooke, H. B. S.) 154–248 (Harvard Univ. Press, 1978).
18. Tobias, P. V. in *Ancestors: The Hard Evidence* (ed. Delson, E.) 94–101 (Liss, New York, 1985).
19. Walker, A. & Leakey, R. E. F. *Sci. Am.* **239**, 44–56 (1978).
20. Stringer, C. B. *J. hum. Evol.* **16**, 135–146 (1987).
21. Stringer, C. B. in *Major Topics in Primate and Human Evolution* (eds Wood, B., Martin, L. & Andrews, P.) 266–294 (Cambridge Univ. Press, UK, 1986).
22. Groves, C. P. & Mazak, V. *Cas. Miner. Geol.* **20**, 225–247 (1975).
23. Chamberlain, A. T. in *Hominidae* (ed. Giacobini, G.) 175–181 (Jaca, Milan, 1989).
24. Wood, B. A. *J. hum. Evol.* **3**, 373–378 (1974).
25. Kennedy, G. E. *J. hum. Evol.* **12**, 587–616 (1983).
26. Wood, B. A. *Nature* **327**, 187–188 (1987).
27. Lewis, O. J. *Functional Morphology of the Evolving Hand and Foot* (Clarendon, Oxford, 1989).
28. Robinson, J. T. *Nature* **240**, 239–240 (1972).
29. Hay, R. L. *Geology of the Olduvai Gorge* (Univ. California Press, Berkeley, 1976).
30. Tobias, P. V. *Nature* **215**, 479–480 (1967).
31. Hill, A., Ward, S., Deino, A., Curtis, G. & Drake, R. *Nature* **355**, 719–722 (1992).
32. Coppens, Y. *Bull. Soc. Geol. Fr.* **21**, 313–330 (1979).
33. Howell, F. C., Haesaerts, P. & de Heinzelin, J. *J. hum. Evol.* **16**, 665–700 (1987).
34. Harris, J. M., Brown, F. H., Leakey, M. G., Walker, A. C. & Leakey, R. E. *Science* **239**, 27–33 (1988).
35. Leakey, M. D. *et al.* *Nature* **262**, 460–466 (1976).
36. Johanson, D. C. & Taieb, M. *Nature* **260**, 293–297 (1976).
37. Johanson, D. C., White, T. D. & Coppens, Y. *Kirtlandia* **28**, 1–14 (1978).
38. Hughes, A. R. & Tobias, P. V. *Nature* **265**, 310–312 (1977).
39. Broom, R. & Robinson, J. T. *Trans. Mus. Mem.* **4**, 11–85 (1950).
40. Clarke, R. J. thesis, Univ. Witwatersrand (1977).
41. Clarke, R. J. & Howell, F. C. *Am. J. Phys. Anthropol.* **37**, 319–336 (1972).
42. Clarke, R. J. *S. Afr. J. Sci.* **73**, 46–49 (1977).
43. Chamberlain, A. T. thesis, Univ. Liverpool (1987).
44. Grine, F. E., Demes, B., Jungers, W. L. & Cole, T. M. (in the press).
45. Tobias, P. V. & von Koenigswald, G. H. R. *Nature* **204**, 515–518 (1964).
46. White, T. D., Johanson, D. C. & Kimbel, W. H. *S. Afr. J. Sci.* **77**, 445–470 (1981).
47. Johanson, D. C. in *Hominidae: Proc. 2nd Int. Congr. Human Paleontology* (ed. Giacobini, G.) 155–166 (Jaca, Milan, 1989).
48. Leakey, R. E. F. *Nature* **248**, 653–656 (1974).
49. Olson, T. R. *J. hum. Evol.* **7**, 159–178 (1978).
50. Leakey, R. E. F. & Walker, A. *Science* **207**, 1103 (1980).
51. Wood, B. A. in *Ancestors: The Hard Evidence* (ed. Delson, E.) 206–214 (Liss, New York, 1985).
52. Groves, C. P. *A Theory of Human and Primate Evolution* (Clarendon, Oxford, 1989).
53. Wood, B. A. *Koobi Fora Research Project IV: Hominid Cranial Remains From Koobi Fora* (Clarendon, Oxford, 1991).
54. Lieberman, D. E., Pilbeam, D. R. & Wood, B. A. *J. hum. Evol.* **17**, 503–511 (1988).
55. Miller, J. A. *Am. J. Phys. Anthropol.* **84**, 385–398 (1991).
56. Wood, B. A., Li, Y. & Willoughby, C. *J. Anat.* **174**, 185–205 (1991).
57. Napier, J. R. *Arch. Biol., Liege* **75** (Suppl.), 673–708 (1964).
58. Susman, R. L. & Stern, J. T. *Science* **217**, 931–934 (1982).
59. Day, M. H. & Napier, J. R. *Nature* **201**, 969–970 (1964).
60. Davis, P. R. *Nature* **201**, 967 (1964).
61. Lewis, O. J. *Functional Morphology of the Evolving Hand and Foot* (Clarendon, Oxford, 1989).
62. Lewis, O. J. *Am. J. Phys. Anthropol.* **37**, 13–34 (1972).
63. Kennedy, G. E. *Am. J. Phys. Anthropol.* **61**, 429–434 (1983).
64. Trinkaus, E. *Am. J. Phys. Anthropol.* **64**, 137–139 (1984).
65. McHenry, H. M. & Corruccini, R. S. *Am. J. Phys. Anthropol.* **49**, 473–488 (1978).
66. Wood, B. A. *Nature* **251**, 135–136 (1974).
67. Leakey, R. E., Walker, A., Ward, C. V. & Grausz, H. M. *Proc. 2nd Int. Congr. Human Paleontology* (ed. Giacobini, G.) 167–173 (Jaca, Milan, 1989).
68. Aiello, L. *J. hum. Evol.* (in the press).
69. Wood, B. A. & Wilson, G. B. in *Variation, Culture and Evolution in African Populations* (eds Singer, R. & Lund, J.) 101–108 (Witwatersrand Univ. Press, Johannesburg, 1986).
70. Hartwig-Schrerer, S. & Martin, R. D. *J. hum. Evol.* **21**, 439–449 (1991).
71. Mayr, E. *Cold Spr. Harb. Symp. quant. Biol.* **15**, 109–118 (1950).
72. Simpson, G. G. *Principles of Animal Taxonomy* (Columbia Univ. Press, New York, 1961).
73. Le Gros Clark, W. E. *The Fossil Evidence for Human Evolution* 1st edn (Univ. Chicago Press, 1955).
74. Robinson, J. T. in *Evolution and Hominisation* 2nd edn (ed. Kurth, G.) 150–175 (Fischer, Stuttgart, 1968).
75. Robinson, J. T. *Nature* **209**, 957–960 (1966).
76. Probst, J. H. *Am. J. Phys. Anthropol.* **52**, 175–189 (1990).
77. Rose, M. D. in *Food Acquisition and Processing in Primates* (eds Chivers, D. J., Wood, B. A. & Bilsborough, A.) 509–524 (Plenum, New York, 1984).
78. Vandevoort, G. in *Evolution des Vertébrés de leur Origine à l'Homme* 450–518 (Masson, Paris, 1969).
79. Eldredge, N. & Tattersall, I. in *Approaches to Primate Paleobiology* (ed. Szalay, F. S.) 218–242 (Karger, Basel, 1975).
80. Wood, B. A. *Cour. Forsch. Inst. Senckenberg* **69**, 99–111 (1984).
81. Delson, E., Eldredge, N. & Tattersall, I. *J. hum. Evol.* **6**, 263–278 (1977).
82. Wood, B. A. & Chamberlain, A. T. in *Major Topics in Primate and Human Evolution* (eds Wood, B., Martin, L. & Andrews, P.) 220–248 (Cambridge Univ. Press, 1986).
83. Bonde, N. in *Major Patterns in Vertebrate Evolution* (eds Hecht, M. K., Goody, P. C. & Hecht, B. M.) 741–804 (Plenum, New York, 1977).
84. Johanson, D. C. & White, T. D. *Science* **202**, 321–330 (1979).
85. Alexeev, V. P. *The Origin of the Human Race* (Progress, Moscow, 1986).
86. Feibel, C. S., Brown, F. H. & McDougall, I. *Am. J. Phys. Anthropol.* **78**, 595–622 (1989).
87. Leakey, R. E., Leakey, M. G. & Behrensmeyer, A. K. (1978).
88. Tobias, P. V. *Palaeont. afr.* **23**, 1–17 (1980).
89. Delson, E. *Nature* **322**, 496–497 (1986).
90. Trinkaus, E. *Am. J. Phys. Anthropol.* **80**, 411–416 (1989).
91. Hartwig-Schrerer, S. *Cour. Forsch. Inst. Senckenberg* (in the press).
92. Falk, D. *Science* **221**, 1072–1074 (1983).
93. Holloway, R. L. in *Hominid Evolution* (ed. Tobias, P. V.) 47–62 (Liss, New York, 1985).
94. Tobias, P. V. *J. hum. Evol.* **16**, 741–761 (1987).

ACKNOWLEDGEMENTS. My researches on the evidence for early *Homo* from Koobi Fora were at the invitation of R. Leakey, and I thank the Director and Trustees of the National Museums of Kenya for their cooperation. I also thank L. Aiello, A. Chamberlain and C. B. Stringer for comments and discussion, and C. Engleman, P. Guest and R. Read for help with the manuscript. This work was supported by The Leverhulme Trust and The Royal Society.

Global mapping of topography on the 660-km discontinuity

Peter M. Shearer & T. Guy Masters

Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, USA

Long-period precursors to the SS seismic phase are used to produce global maps of topography for the discontinuity at 660 km depth in the upper mantle. These maps indicate discontinuity depth variations of up to 30 km and suggest a correlation between regional depressions in the 660-km discontinuity and subduction zones — a result more consistent with models in which the subducting slabs are deflected horizontally at the discontinuity than with models of unhindered slab penetration into the lower mantle.

MANY of the fundamental unresolved questions in mantle chemistry and dynamics hinge on the properties of the Earth's 660-km discontinuity in seismic velocities. Advocates of whole-mantle convection suggest that the discontinuity represents a mineral phase change alone which causes little or no resistance to convection, whereas others favour a change in composition near 660 km which separates the mantle into two convection regimes. For the whole-mantle model, downgoing slabs in subduction zones penetrate through the 660-km discontinuity into the lower mantle; in the layered-mantle model the slabs encounter resistance near 660 km which prevents penetration. Intermediate models have also been proposed in which the 660-km discontinuity is a partial barrier to convection^{1,2}. Detailed knowledge of any depth variations of the 660-km discontinuity would help to discriminate between these models, as they predict different patterns of discontinuity topography.

The existence of a sharp jump in seismic velocities near 660 km depth has been known for over 25 years^{3,4}. High-pressure experiments in mineral physics have shown that the bulk of this velocity jump can be explained by the phase change γ -olivine (spinel) $\rightarrow$ perovskite + magnesiowüstite⁵⁻⁷. But controversy remains⁸⁻¹⁰ about whether a change in composition might also occur at or near 660 km. The presence of a compositional density contrast would inhibit convection across the boundary. Even without such a change in chemistry, slabs may encounter some resistance near 660 km due to a viscosity contrast at the discontinuity¹¹ and the negative Clapeyron slope (the change in pressure with temperature, $\gamma_{660} = dP/dT$) for the 660-km phase change^{2,12}. Values for γ_{660} obtained from mineral physics experiments can be used to relate variations in mantle temperature to differences in discontinuity depths; colder temperatures should cause the discontinuity to move to deeper depths.

In principle, seismic travel times can be used to obtain the velocity structure near where the subducting slabs intersect the 660-km discontinuity in order to resolve whether the slabs penetrate to the lower mantle. But results have been inconsistent, with some studies finding evidence for deep slab penetration¹³⁻¹⁸, whereas others¹⁹⁻²¹ seem to image the horizontal deflection of slabs in some areas near 660 km. Another approach is to study directly the behaviour of the 660-km discontinuity near the subducting slabs by examining reflected and converted phases from the discontinuity^{22,23}. In practice this has been hindered by the difficulty of observing these relatively low-amplitude seismic arrivals. Recently, however, processing tech-

niques have been developed that permit these arrivals to be seen fairly routinely in long-period seismograms^{22,24-26}.

Here we examine long-period precursors to SS which result from underside reflections from the 660-km discontinuity. These S660S phases are ideally suited for studying global variations in discontinuity topography because of their wide distribution of bounce points. Previous analysis of S660S arrivals²⁶ showed that there are only small differences in discontinuity depths between different tectonic regions and suggested the presence of a 20-km depression in the vicinity of the subduction zones in the northwest Pacific. We now extend this analysis to a larger data set and directly map the 660-km discontinuity topography without dividing the data into different tectonic regions. Maps obtained from analysis of more than 3,000 seismograms indicate large-scale coherent patterns of discontinuity topography with depth variations of up to 30 km. Regional depressions in the 660-km discontinuity seem to be associated with subduction zones, with the position and broad extent of these depressions supporting models in which the subducting slabs break up or are deflected horizontally at 660 km rather than models of unhindered slab penetration.

Precursors to SS

We examine long-period transverse-component seismograms at source-receiver ranges from 80° to 180° obtained from the Global Digital Seismograph Network between 1976 and 1987 and select more than 4,000 traces with high-quality SS arrivals from shallow events (<75 km deep). Figure 1a shows the result of stacking these waveforms after normalizing and aligning the SS arrival at zero time as a reference phase. Amplitudes in the stacked image are indicated by colours; the maximum amplitude shown is only 0.05 of the SS wave amplitude. Figure 1b shows the predicted arrival times of seismic phases relative to SS as obtained by ray tracing through a reference model of the Earth²⁷. The prominent blue streaks seen 2.5 to 4 min before SS are precursors resulting from underside reflections off upper-mantle discontinuities. The reflections from the 410- and 660-km discontinuities, S410S and S660S, are identified in Fig. 1b. The stacked image also shows an apparent arrival from a reflector between 410 and 660 km, corresponding to a discontinuity at 520 km depth²⁵. The other discontinuity phases labelled in Fig. 1b are only marginally visible in Fig. 1a.

Here we consider only the S660S phase; some results for S410S and other discontinuity phases have been described elsewhere²⁶. Figure 2 shows the ray geometry of S660S compared with S and SS ray paths. S660S is ideally suited for examining discontinuity topography as it is associated with a single reflection point on the 660-km discontinuity. In contrast, most other discontinuity phases result from several different possible ray paths and reflection points, greatly complicating their interpretation. Figure 3 shows S660S bounce points from 3,139 seismograms between 110° and 180°. Although some areas of sparse sampling exist, reasonable global coverage is obtained, with the Pacific and Indian Oceans being particularly well sampled. To find possible S660S phases on individual seismograms, we calculate the cross-correlation function between the SS arrival and the remainder of the seismogram (details of this method are described in ref. 26). This procedure obtains travel times and amplitudes of apparent S660S arrivals relative to the SS travel

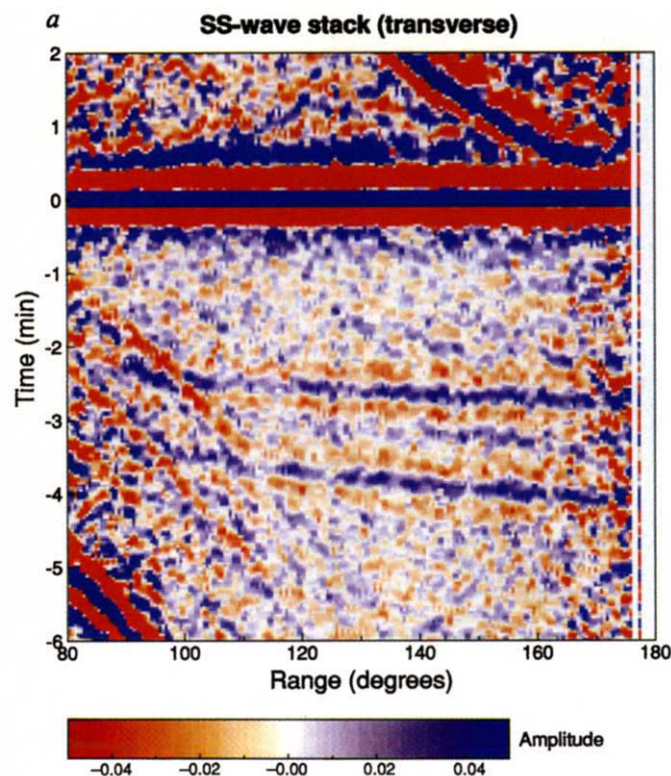


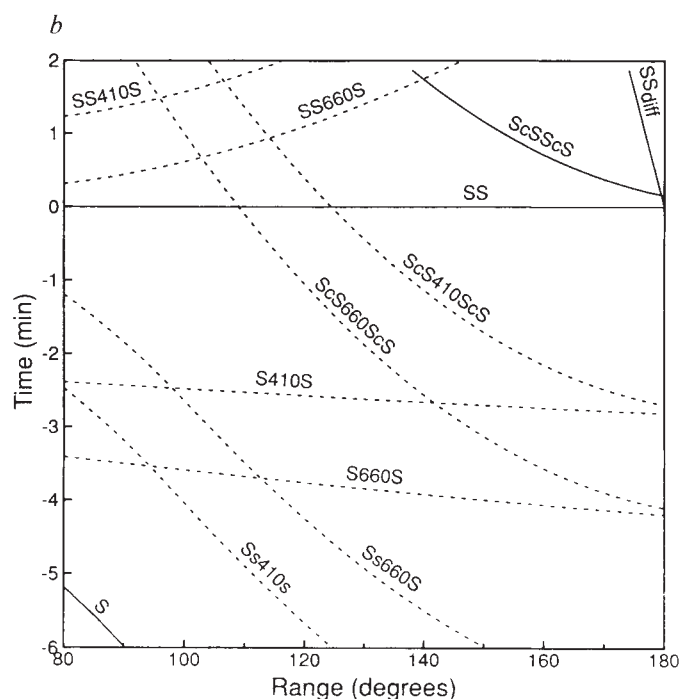
FIG. 1 a, Stacked colour image of more than 4,000 long-period transverse-component seismograms from the Global Digital Seismic Network, compared

time. These SS – S660S differential travel times are directly sensitive to the two-way travel time between the discontinuity and the surface (see Fig. 2), and, assuming an upper-mantle velocity model, can be converted to apparent discontinuity depths.

Discontinuity depths measured in this way will be biased by any lateral velocity variations in the upper mantle, and must be corrected. For example, a slow region above the 660-km discontinuity would increase the SS – S660S travel time and increase the apparent depth of the 660-km discontinuity. A direct measure of vertical travel time differences through the upper mantle may be obtained from SS – S times. These phases have nearly identical ray paths near the source and receiver, but SS has two additional legs through the upper mantle near the SS bounce point (see Fig. 2). Because recent global velocity inversions have indicated that heterogeneity in the mantle below 660 km is relatively small compared with differences above 660 km, the SS – S travel times provide a good estimate of the upper mantle velocity variations at the SS bounce point. We use a smoothed version of SS – S residuals obtained by Woodward and Masters²⁸ to correct our observed SS – S660S residuals. Analysis of the 410-km phases is complicated by the effects of transition-zone heterogeneity and is deferred to future work.

Topography of the 660-km discontinuity

In principle, 660-km discontinuity depths could be obtained for each of the S660S bounce points shown in Fig. 3. Because of the large scatter in observed arrival times, however, stable estimates of discontinuity depths are only possible when results are averaged from a number of seismograms. Figure 4a shows the variations in 660-km discontinuity depths (relative to a mean depth of 659 km) obtained for 416 caps spaced $\sim 10^\circ$ apart. Within each cap, results are combined from all S660S bounce points within a 10° radius, and the mean apparent depth to the 660-km discontinuity and its standard error are calculated with a bootstrap technique²⁹. Only caps with standard errors in depth less than 10 km are shown in Fig. 4a, with the gaps in the spatial coverage generally corresponding to areas with little or no data.



with (b) travel-time curves for primary seismic phases (solid lines) and upper-mantle discontinuity phases (dashed). Positive amplitudes are shown in blue, negative amplitudes in red, with the scale ranging up to 0.05 of the maximum amplitude of the SS reference phase. We examine the S660S phase, which arrives 3.5–4 min before SS and results from an underside reflection off the 660-km discontinuity.

Large-scale coherent patterns of discontinuity topography are seen with peak-to-peak amplitudes of ~ 30 km.

The depth variations shown in Fig. 4a correspond to differences in the relative arrival times of S660S and SS. Conceivably these times could be biased by effects other than discontinuity topography. Probably the largest potential source of error is inaccuracies in the corrections for lateral heterogeneity, as these will directly affect the apparent depths that we calculate. The gross structures shown in Fig. 4 are most reliably resolved, because the SS – S residual pattern used for the corrections is well constrained at large scales. Improved velocity models will, however, almost certainly result in some changes to the finer details seen in Fig. 4. Another possibility is that a small discontinuity of fluctuating brightness or depth exists near the 660-km discontinuity, biasing S660S travel times by variable amounts. If so, one might expect to see some correlation between apparent discontinuity depths and the amplitude of discontinuity reflections. Amplitudes of the S660S phases presented here are not correlated to the apparent depths to the 660-km discontinuity, although such a correlation has been noted in ScS reverberation data²⁴.

Using the method of spherical splines^{28,30}, we used these cap-averaged depths and their standard errors to produce a smooth model of discontinuity topography (Fig. 4b). This technique gives the surface with smallest second lateral derivative integrated over the sphere for a given target χ^2 . The model

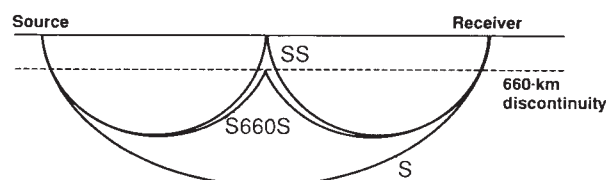


FIG. 2 The ray geometry for S, SS and S660S. The ray paths of SS and S660S are nearly identical except that the S660S bounce point is at the 660-km discontinuity rather than the surface.

TABLE 1 Spherical harmonic coefficients for model Topo660a

	$m=0$	1	2	3	4	5	6	7	8
C_0^m	658.537								
S_0^m	0.000								
C_1^m	1.036	-0.691							
S_1^m	0.000	-0.922							
C_2^m	-0.973	0.225	0.896						
S_2^m	0.000	2.506	-1.731						
C_3^m	0.555	0.225	0.258	-0.011					
S_3^m	0.000	1.981	0.688	0.494					
C_4^m	-0.204	-1.217	-0.022	0.735	-0.458				
S_4^m	0.000	-0.783	-0.901	0.790	-0.426				
C_5^m	0.806	-0.278	0.651	0.077	-0.265	-0.254			
S_5^m	0.000	0.172	-0.617	-0.115	-0.250	-0.317			
C_6^m	-0.153	-0.241	0.189	0.199	0.035	0.281	0.300		
S_6^m	0.000	-0.225	-0.386	0.277	-0.189	-0.546	0.349		
C_7^m	0.141	-0.017	0.285	0.194	-0.160	0.157	-0.112	-0.116	
S_7^m	0.000	0.116	0.082	-0.042	-0.439	-0.374	0.328	0.116	
C_8^m	-0.157	0.015	-0.130	0.266	-0.047	-0.246	0.029	-0.179	0.162
S_8^m	0.000	-0.060	0.124	0.117	0.358	0.268	0.240	-0.178	-0.074

Spherical harmonic coefficients up to degree 8 for the model of discontinuity topography shown in Fig. 4b. We have normalized the coefficients using the convention described in ref. 56.

shown fits the 340 cap-averaged depths with a χ^2 of 300. Changing the target χ^2 so that the data are substantially over- or underfitted results in maps of similar shape to that in Fig. 4b but with variable amplitude. The equivalent spherical harmonic expansion of this surface has nearly all its power in the first eight harmonic degrees (Table 1) and is dominated by harmonic degree 2. Unlike simple spherical harmonic expansions, the spherical spline method does not produce models with large structure in regions with no data. The topography shown in Fig. 4b should not, however, be taken at face value without comparison with Fig. 4a to see where the data actually constrain the model.

The 660-km discontinuity is anomalously deep in regions northwest of the Japan and Kuril-Kamchatka trenches, west of the Tonga trench, and in eastern South America. Depths are generally shallow beneath North America, Antarctica and the Indian Ocean. The data coverage in the northwest Pacific is especially good and clearly resolves a 15–20-km-deep trough in the 660-km discontinuity which is ~1,500 km wide and 5,000 km long. The corrections for upper-mantle heterogeneity from the SS–S times are relatively small in this region; this feature is also apparent in the uncorrected data. This trough is oriented parallel to the Kuril-Kamchatka subduction zone but is displaced to the northwest. The depressions west of Tonga and in eastern South America are also near subduction zones, but should be considered less reliable anomalies, as the Tonga feature is constrained by comparatively little data and the South American depression is largely determined by the corrections for fast upper-mantle velocities in this area. It is possible that additional topographic lows occur for other subduction zones in the western Pacific, but the data coverage is too sparse to resolve discontinuity depths in this region.

Transition zone temperature variations

Large-scale topographic variations on the 660-km discontinuity seem to be about ~30 km in peak-to-peak amplitude, consistent with results from other discontinuity phases and ScS reverberations. Measured values of the Clapeyron slope for the phase change γ -olivine (spinel) $\rightarrow$ perovskite + magnesiowüstite near 660 km have ranged from -2 to -4 MPa K⁻¹ (refs 31–33). The negative Clapeyron slopes imply an endothermic reaction in which the discontinuity moves downwards at lower temperatures. These values predict that every 100-K decrease in mantle temperature will cause a 5–10-km increase in depth to the 660-km discontinuity. Thus, lateral differences in mantle temperature at 660 km between 300° and 600° are required to explain the observed 30 km of discontinuity topography directly

in terms of the Clapeyron slope, with the topographic highs and lows representing warmer and cooler regions respectively.

Because temperature variations affect seismic velocities, one might expect these features to show up in global inversions for transition-zone velocity structure. For some time, these inversions^{34–37} have indicated the presence of a large anomaly of spherical harmonic degree 2 in the transition zone, with fast velocities centred in the western Pacific and eastern South America. The latter feature is roughly consistent with the 660-km depression seen in the maps of discontinuity topography (Fig. 4), but the western Pacific anomaly (centred east of the Philippines) is not reproduced in these maps. Few data exist in this area, so it is unlikely that a significant depression in this area could be resolved with the S660S phases considered here. ScS reverberation data²⁴ have much better coverage in this region, but also do not show a significant depression in the 660-km discontinuity.

The most clearly resolved feature in the S660S data is the trough in the 660-km discontinuity behind the Kuril-Kamchatka subduction zone. Global models of transition-zone velocity do not show anomalies specific to this region, although in some models the western Pacific high-velocity zone extends to the north of Japan. The high in the 660-km topography near Antarctica is in reasonable agreement with a slow region in the velocity models, but highs over the Indian Ocean and northern North America do not seem correlated with velocity features. Comparisons between discontinuity topography and mantle velocity models should be made with care, as the models have different vertical and horizontal resolution. The global velocity inversions

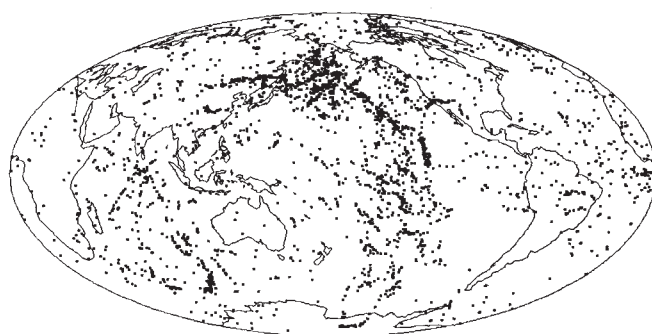


FIG. 3 The global distribution of S660S bounce points at source-receiver ranges between 110° and 180°. Estimates of 660-km discontinuity depths can be made at each of these points, but in practice stable results are obtained only by averaging results from many points.

have limited depth resolution and give only average anomalies within a broad depth range, whereas variations in 660-km discontinuity depth are sensitive only to the temperature at the discontinuity itself. In addition, the models are not derived independently, as knowledge of upper-mantle velocities is necessary to correct apparent discontinuity depths for the integrated effect of heterogeneity between the surface and 660 km.

Northwest Pacific subduction zones

The data coverage near the subducting slabs in the northwest Pacific is dense enough for a more detailed examination of discontinuity topography in this area. Figure 5 is a close-up of this region and shows the location of the S660S bounce points and the subduction zone seismicity. Earthquake locations from the International Seismological Centre (ISC) from 1970 to 1987 are shown, with crosses for events between 100 and 400 km depth and squares for events deeper than 400 km. These earthquakes define the Wadati-Benioff zones which outline the position of the subducting slabs. The line labelled AB defines a great-circle cross-section through this region; the adjacent lines are 5° away. These lines are chosen to enclose a significant number of the S660S bounce points rather than to be exactly orthogonal to the Kuril-Kamchatka subduction zone.

S660S data are binned by range along AB, and discontinuity depth estimates obtained for each range bin. Figure 6 shows the resulting 660-km discontinuity depths and standard error bars for both the raw and SS-S corrected data. Depths are shown at 3° range increments with each point including data within a window of 6° range. A 20-km-deep depression is seen in the 660-km discontinuity, consistent with the position of the trough shown in Fig. 4. This feature is also apparent in the uncorrected depths, but its shape is distorted by a gradient in integrated upper-mantle velocities along the cross-section. The top plot also shows ISC earthquake locations (1970-87) within 3° of AB. The event at 611 km is the 12 May 1990 Sakhalin Island earthquake³⁸. Because the cross-section crosses the Kuril-Kamchatka subduction zone rather obliquely, some broadening is seen in the Wadati-Benioff zone, which appears nearly vertical

because of the large vertical exaggeration. In addition, this profile is near the corner between the Kuril and Japan subduction zones, complicating any interpretation of slab geometries. It is clear, however, that the bulk of the depression in the 660-km discontinuity is offset to the northwest from the point at which the projection of the subducting slabs intersects the discontinuity.

The depression appears to be ~15° (1,700 km) across and ~20 km deep. The lateral resolution of the long-period S660S phase (~25-s period) is limited to features larger than ~10°. In particular, these data cannot exclude the possibility of a large deflection in the 660-km discontinuity in the immediate vicinity of subducting slabs if this feature were only a few degrees across. The edge of the trough could be fairly sharp, particularly on the southeastern side (note that this sharpness is lost in the smooth map of topography shown in Fig. 4). There is a suggestion in the cross-section (Fig. 6) that the 660-km discontinuity may be slightly higher in areas immediately adjacent to the trough than regions further away. But, considering the size of the error bars and the absence of this feature in the cap averages (Fig. 4), the existence of these highs is very uncertain.

Slab penetration?

The idea that the subducting slabs in the northwest Pacific penetrate the 660-km discontinuity is controversial. Residual sphere analyses of both S-waves and P-waves for deep focus events have revealed patterns consistent with penetration of the Kuril slab to depths of 900 km or more^{13,14}, but it has been argued that these patterns are caused by deep-mantle or near-receiver anomalies unrelated to the slab structure³⁹⁻⁴². Waveform and amplitude anomalies have also been used to infer deep slab penetration in the Kurils^{18,43}, but these interpretations have been questioned^{41,44,45}. Three-dimensional P-wave velocity inversions from ISC travel times for structure in this region have found evidence both for deep slab penetration¹⁷ and the horizontal deflection of slabs near 660 km (ref. 19). A study that used both direct P and surface-reflected pP travel times found evidence for both types of structures, with deep slab anomalies imaged in the northern Kurils and near-horizontal features seen in the Japan and Izu-Bonin arcs²⁰. Recent work combining whole-mantle and regional inversion for western Pacific structure²¹ appears to resolve horizontal slab features west of the southern Kuril to Izu-Bonin arcs which extend more than 1,000 km. The regional depression in the 660-km discontinuity behind the Kuril subduction zone (Figs 4 and 6) supports those models in which the cold, subducting slab spreads out horizontally near 660 km rather than models of unhindered slab penetration.

A similar debate has occurred over the fate of subducting slabs in the Tonga region. The complex geometry and focal mechanisms of deep earthquakes from the Tonga subduction zone may indicate that the slabs encounter resistance near 660 km and are displaced horizontally^{46,47}. Alternatively, the contortions in the Tonga slab may be due to local tectonic

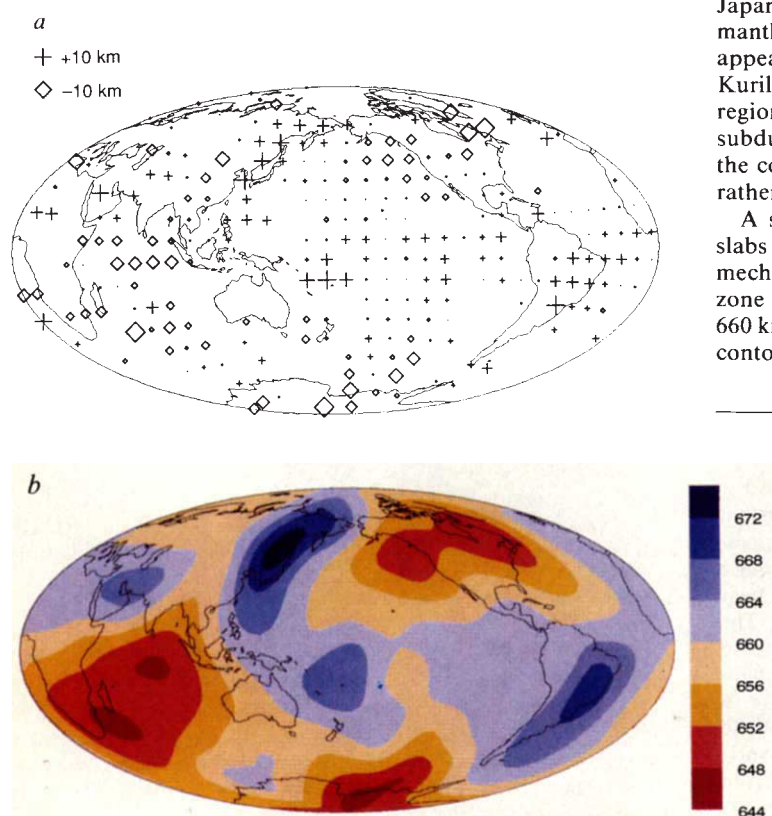


FIG. 4 *a*, Average depths to the 660-km discontinuity obtained for data within caps of 10° radius (relative to a mean depth of 659 km). Only cap averages with standard errors less than 10 km are plotted. *b*, Smooth map of topography on the 660-km discontinuity obtained using the method of spherical splines on the cap averages shown in *a*. Peak-to-peak topography variations are ~30 km. Note the depressions in the 660-km discontinuity north of Japan, west of Tonga, and in eastern South America.

FIG. 5 S660S bounce points and seismicity in the north-western Pacific and eastern Asia (■, S660S bounce points; +, earthquakes between 100 and 400 km; □, earthquakes deeper than 400 km). The curve labelled AB is a great circle that defines the cross-section shown in Fig. 6; the adjacent curves are 5° away.

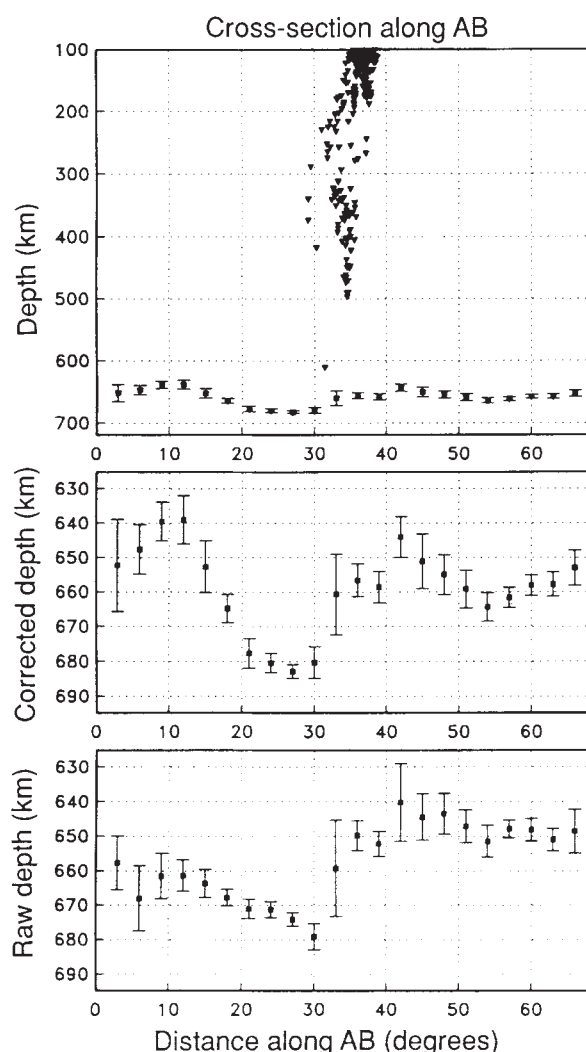
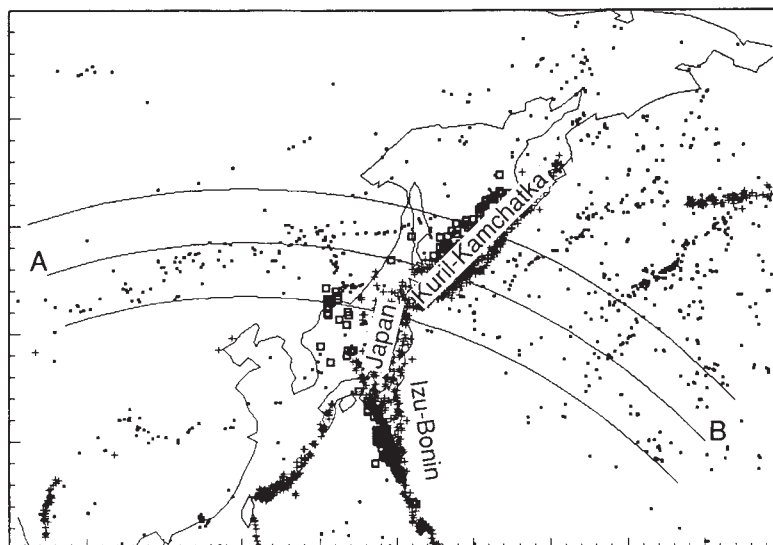


FIG. 6 Depths to the 660-km discontinuity and seismicity along a cross-section through the Kuril-Kamchatka subduction zone (the line AB in Fig. 5), showing results from both raw SS-S660S times (bottom plot) and depths corrected for the effects of upper-mantle heterogeneity (top plots). Note the depression, ~15° wide, in the 660-km discontinuity just behind where the projection of the subducting slab seems to intersect the discontinuity.

interactions and may not necessarily require resistance to penetration at 660 km (ref. 48). Some travel-time studies^{16,49-51} from deep-focus events have found evidence for high-velocity anomalies consistent with deep slab penetration whereas others have not^{52,53}. The suggestion of a regional depression in the 660-km discontinuity behind the Tonga subduction zone (Fig. 4) supports the idea that at least some of the subducted lithosphere has been deposited at this depth. Evidence for this depression is not conclusive, as it is constrained by only a few S660S phases and the SS-S corrections for upper-mantle structure are also sparse in this area. Another depression in the 660-km discontinuity is seen in eastern South America, and may be related to subduction beneath the Andes. This feature is not clear in the raw S660S times but results from the corrections for the fast upper-mantle velocities in this region.

If the 660-km discontinuity is entirely a compositional change, modelling results have shown that it should be displaced several hundred kilometres or more by subducting slabs^{12,54}. Our S660S results and other recent observations^{23,24} are clearly inconsistent with such large depth variations, and support the idea that most of the jump in seismic velocities near 660 km results from a phase change. The possibility of a small chemical change in density at or near 660 km cannot be ruled out, because even large perturbations to its depth might not be seen in seismic observations. Even in the absence of a compositional density change at 660 km, the subducting slabs will encounter some resistance to penetration because of the negative Clapeyron slope of the 660-km phase change and the likely existence of a viscosity contrast near 660 km (ref. 11). Recent modelling results² have indicated that a Clapeyron slope of $\gamma_{660} = -4 \text{ MPa K}^{-1}$ will result in layered convection, with a mixture of whole-mantle and layered convection occurring for $\gamma_{660} = -2 \text{ MPa K}^{-1}$.

Our results suggest a correlation between subduction zones and broad regional depressions in the 660-km discontinuity. The existence of these depressions is consistent with models in which the descending slabs in many areas are deflected horizontally to create large areas of moderately cooler temperatures behind subduction zones. Such models would include the hypothesis of a gravitationally trapped megalith⁵⁵ and some numerical simulations of subducting slabs encountering a viscosity contrast¹¹. The possibility of localized regions of deep slab penetration, as suggested by some velocity inversions, cannot be excluded. In addition, the data are currently insufficient to determine whether these discontinuity depressions are a general feature beneath all deep subduction zones, or are specific to certain regions. □

Received 24 October 1991; accepted 24 January 1992.

1. Silver, P. G., Carlson, R. W. & Olson, P. *Ann. Rev. Earth planet. Sci.* **16**, 477–541 (1988).
2. Machetel, P. & Weber, P. *Nature* **350**, 55–57 (1991).
3. Niazi, M. & Anderson, D. L. *J. geophys. Res.* **70**, 4633–4640 (1965).
4. Johnson, L. R. *J. geophys. Res.* **72**, 6309–6325 (1967).
5. Liu, L.-G. *Nature* **262**, 770–772 (1976).
6. Jackson, I. *Earth planet. Sci. Lett.* **62**, 91–103 (1983).
7. Ito, E., Takahashi, E. & Matsui, Y. *Phys. Earth planet. Inter.* **67**, 238–248 (1984).
8. Jeanloz, R. in *Mantle Convection: Plate Tectonics and Global Dynamics* (ed. Peltier, W. R.) 203–259 (Gordon and Breach, New York, 1989).
9. Duffy, T. S. & Anderson, D. L. *J. geophys. Res.* **94**, 1895–1912 (1989).
10. Jeanloz, R. *Geophys. Res. Lett.* **18**, 1743–1746 (1991).
11. Gurnis, M. & Hager, B. H. *Nature* **335**, 317–321 (1988).
12. Christensen, U. R. & Yuen, D. A. *J. geophys. Res.* **89**, 4389–4402 (1984).
13. Jordan, T. H. *J. geophys. Res.* **43**, 473–496 (1977).
14. Creager, K. C. & Jordan, T. H. *J. geophys. Res.* **89**, 3031–3049 (1984).
15. Creager, K. C. & Jordan, T. H. *J. geophys. Res.* **91**, 3573–3589 (1986).
16. Fischer, K. M., Creager, K. C. & Jordan, T. H. *J. geophys. Res.* **96**, 14403–14427 (1991).
17. Kimaya, S., Miyatake T. & Hirahara, K. *Geophys. Res. Lett.* **15**, 829–831 (1988).
18. Suetsugu, D. *J. Phys. Earth* **37**, 265–295 (1989).
19. Zhou, H. & Clayton, R. W. *J. geophys. Res.* **95**, 6829–6851 (1990).
20. van der Hilst, R., Engdahl, R., Spakman, W. & Nolet, G. *Nature* **353**, 37–43 (1991).
21. Fukao, Y., Obayashi, M., Inoue, H. & Nenbai, M. *J. geophys. Res.* (in the press).
22. Revenaugh, J. & Jordan, T. H. *J. geophys. Res.* **94**, 5787–5813 (1989).
23. Richards, M. A. & Wicks, C. W. *Geophys. J. Int.* **101**, 1–35 (1990).
24. Revenaugh, J. & Jordan, T. H. *J. geophys. Res.* **96**, 19763–19780 (1991).
25. Shearer, P. M. *Nature* **344**, 121–126 (1990).
26. Shearer, P. M. *J. geophys. Res.* **96**, 18147–18182 (1991).
27. Dziewonski, A. M. & Anderson, D. L. *Phys. Earth planet. Inter.* **25**, 297–356 (1981).
28. Woodward, R. L. & Masters, G. *J. geophys. Res.* **96**, 6351–6377 (1991).
29. Etrou, B. & Tibshirani, R. *Science* **253**, 390–395 (1991).
30. Shure, L., Parker, R. L. & Backus, G. E. *Phys. Earth planet. Inter.* **28**, 215–229 (1982).
31. Ito, E. & Yamada, H. *High-Pressure Research in Mineral Physics* (ed. Akimoto, S. & Manghnani, M. H.) 405–419 (Center for Academic Publications, Tokyo, 1982).
32. Ito, E. & Takahashi, E. *J. geophys. Res.* **94**, 10637–10646 (1989).
33. Ito, E., Akaogi, M., Topor, L. & Navrotsky, A. *Science* **249**, 1275–1278 (1990).
34. Masters, G., Jordan, T. H., Silver, P. G. & Gilbert, F. *Nature* **298**, 609–613 (1982).
35. Woodhouse, J. H. & Dziewonski, A. M. *J. geophys. Res.* **89**, 5953–5986 (1984).
36. Tanimoto, T. *Geophys. J. Int.* **100**, 327–336 (1990).
37. Masters, G. & Bolton, H. *Eos* **72**, 316 (1991).
38. Ekström, G., Dziewonski, A. M. & Ibañez, J. *Eos* **71**, 1462 (1990).
39. Zhou, H., Anderson, D. L. & Clayton, R. W. *J. geophys. Res.* **95**, 6799–6827 (1990).
40. Schwartz, S. Y., Lay, T. & Grand, S. P. *Geophys. Res. Lett.* **18**, 1265–1268 (1991).
41. Schwartz, S. Y., Lay, T. & Beck, S. L. *J. geophys. Res.* **96**, 14445–14460 (1991).
42. Gaherty, J. B., Lay, T. & Vidale, J. E. *J. geophys. Res.* **96**, 16349–16367 (1991).
43. Silver, P. & Chan, W. *J. geophys. Res.* **91**, 13787–13802 (1986).
44. Beck, S. L. & Lay, T. *Geophys. Res. Lett.* **13**, 1007–1010 (1986).
45. Lay, T. & Young, C. J. *Geophys. Res. Lett.* **16**, 605–608 (1989).
46. Giardini, D. & Woodhouse, J. H. *Nature* **307**, 505–509 (1984).
47. Giardini, D. & Woodhouse, J. H. *Nature* **319**, 551–555 (1986).
48. Hamburger, M. W. & Isacks, B. L. *J. geophys. Res.* **92**, 13841–13854 (1987).
49. Fitch, T. J. *Phys. Earth planet. Inter.* **26**, 156–166 (1975).
50. Frohlich, C. & Barazangi, M. *Phys. Earth planet. Inter.* **21**, 1–14 (1980).
51. Zhou, H. *Phys. Earth planet. Inter.* **61**, 199–229 (1990).
52. Sondergeld, C. H., Isacks, B. L., Barazangi, M. & Billington, S. *Bull. seismol. Soc. Am.* **67**, 537–541 (1977).
53. Bock, G. *Phys. Earth planet. Inter.* **25**, 360–371 (1981).
54. Hager, B. H. & Raefsky, A. *Eos* **62**, 1074 (1981).
55. Ringwood, A. E. & Irfune, T. *Nature* **331**, 131–136 (1988).
56. Stacey, F. D. *Physics of the Earth* (Wiley, New York, 1977).

ACKNOWLEDGEMENTS. We thank S. King, J. Phipps Morgan and R. Woodward for advice. This research was supported in part by the NSF.

Nature of biological electron transfer

Christopher C. Moser, Jonathan M. Keske, Kurt Warncke, Ramy S. Farid & P. Leslie Dutton*

The Johnson Research Foundation, Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Powerful first-order analysis of intraprotein electron transfer is developed from electron-transfer measurements both in biological and in chemical systems. A variation of 20 Å in the distance between donors and acceptors in protein changes the electron-transfer rate by 10^{12} -fold. Protein presents a uniform electronic barrier to electron tunnelling and a uniform nuclear characteristic frequency, properties similar to an organic glass. Selection of distance, free energy and reorganization energy are sufficient to define rate and directional specificity of biological electron transfer, meeting physiological requirements in diverse systems.

IN 1960 Chance and Nishimura discovered that electron transfer between a cytochrome *c* and a light-activated bacteriochlorophyll in the bacterium *Chromatium vinosum* occurred even at 80 K (ref. 1). Later DeVault and Chance² reported that the millisecond electron transfer was temperature-independent from 120 K to 4 K and deduced that the reaction proceeded through a quantum mechanical tunnelling mechanism, perhaps taking place over a distance as large as 30 Å. They suggested that such long-range electron transfer may be common and essential to the biological redox energetics of chloroplasts and mitochondria. This work coincided with Mitchell's development of the chemiosmotic hypothesis which proposed that the key electron-transfer steps of respiration and photosynthesis were directed across the 35 Å span of the membrane dielectric³. The past 25

years have seen these advances gain support and general acceptance, yet descriptions of factors that govern long-range electron transfer vital to the bioenergetic machinery have remained unclarified. The analysis we present here provides guidelines basic to the understanding of the design and engineering of respiratory and photosynthetic electron-transfer chains and other redox enzymes.

The traditional transition-state description of the rates of chemical reactions involves motion along a potential energy surface in which the reactant atoms gain energy from thermal collisions, surmount an activation energy barrier to achieve a transition state and spontaneously decay into the product. In contrast to these essentially adiabatic reactions, where formation of the transition state leads almost inevitably to the product, the probability of long-distance electron transfer between distant, weakly coupled donors and acceptors from such a transition point is small. Instead, a simple, non-adiabatic description given by Fermi's golden rule is more appropriate (equations and parameters are described in Box 1 and ref. 4). The rule states that the electron-transfer rate is proportional to the square of the (weak) coupling of the reactant and product electronic states, V_R^2 , which in turn is proportional to the overlap of the donor and acceptor electronic wavefunctions across the space separating electron donor and acceptor molecules. In the simplest view, the overlap falls off exponentially with edge-to-edge distance (R), modified by a coefficient (β) which describes the contribution made by the intervening medium in propagating the wavefunction.

The nuclear positions of reacting molecules and their environment, so critical to transition state theory, have a central role in the second term of the golden rule. Calculation of this term, described as the density of states by Fermi, is problematic but can be considerably simplified by assuming the nuclei of the donor, acceptor and immediate environment act effectively as harmonic oscillators. In this case, a Franck-Condon-weighted density of states reflecting the overlap of reactant and product

* To whom correspondence should be addressed.

oscillators can be written explicitly. The classical view of nuclear motion described by Marcus^{5,6} leads to equation (3a) of Table 1, whereas a quantum view of the harmonic oscillators leads to the quantum-mechanically correct expressions of Levich and Dogonadze⁷ and of Jortner⁸ (equation (3b)) and the semi-classical expression of Hopfield⁹ (equation (3c)).

Figure 1 illustrates reactant and product harmonic oscillator potential surfaces and the Franck-Condon factors derived from the overlap of the nuclear wavefunctions. The overlap of reactant and product vibrational states will be maximal when $-\Delta G^\circ$ (where ΔG° is the standard Gibbs free energy) matches the reorganization energy (λ), defined as the energy required to distort the equilibrium nuclear geometry of the reactant state into the equilibrium geometry of the product state without electron transfer. Experiment has verified an overall roughly gaussian dependence of electron-transfer rate on $-\Delta G^\circ$ predicted by Marcus^{4,5}. Thus the rate rises with increasing $-\Delta G^\circ$, reaches a maximum when $-\Delta G^\circ = \lambda$, and falls with $-\Delta G^\circ$ values greater than λ . In the classical view, the breadth of the gaussian is proportional to $\sqrt{2\lambda kT}$, where kT is the Boltzmann thermal energy. But the quantum view is required when the effective frequency (ω) of the oscillators is large and $\hbar\omega$ is greater than $2kT$. Under these conditions, tunnelling of the nuclei as well as the electrons is important and the breadth of the gaussian curve becomes relatively temperature-independent and proportional to $\sqrt{\lambda \hbar\omega}$. Thus, temperature becomes an important experimental variable in estimating $\hbar\omega$.

Application of theory to an increasing number of systems is now possible. Here we have attempted to draw from theory a single group of measured parameters that can be applied commonly to systems in chemistry and biology. Hence, we restrict our attention to electron transfers in which the distance is known either from X-ray analysis or, in the case of rigid covalent systems, from molecular modelling to an uncertainty of $<2 \text{ \AA}$. We also required in the first instance that electron-transfer rates were measured over a suitable range of free energies measured in the same or similar medium. Recent studies on the bacterial photosynthetic reaction centre (Fig. 2a) have provided a particularly rich source of thermodynamic, kinetic and structural information on intramolecular electron transfer^{6,10-13}. Analogous studies have explored intramolecular electron-transfer reactions in semisynthetic^{14,15} and fully synthetic systems¹⁶⁻¹⁹. Several representative structures are shown in Fig. 2b-e.

For each of these systems we first examine the Franck-Condon factors, including the effects of $-\Delta G^\circ$, T , $\hbar\omega$ and λ in modulating the electron-transfer rates, and obtain the optimal rates for each reaction. The optimal rate obtained when $-\Delta G^\circ$ matches λ provides the vantage point for viewing the contributions to the rate from the electronic terms because here the rate is least affected by variations in $\hbar\omega$, λ and T . Thus, comparing the optimal rates in the variety of reactions at different distances permits evaluation of the electronic terms and the contributions of V_R and β .

Figure 3a (top) illustrates the $-\Delta G^\circ$ dependence at 298 K of the rate of the physiologically productive, $10 \text{ \AA}$ electron transfer from the photo-reduced bacteriopheophytin, BPh^- , to various quinones replacing the native quinone of the Q_A site in reaction centres from *Rhodobacter sphaeroides*¹³. The same reaction studied at smaller $-\Delta G^\circ$ values in *Rhodopseudomonas viridis* (distance $9.6 \text{ \AA}$) extends the trend, indicating that the reaction is independent of the species (J.M.K. *et al.*, manuscript in preparation). Figure 3a (bottom) shows a similar $-\Delta G^\circ$ dependence, shown here at 35 K, for the physiologically non-productive electron transfer that occurs in *Rb. sphaeroides* over a $22.5 \text{ \AA}$ from the Q_A^- back to the oxidized bacteriochlorophyll dimer, BChl_2^+ . A scatter in the rates of the order of 10-fold is typical of virtually all measurements in which the $-\Delta G^\circ$ is changed chemically. Despite this uncertainty, several conclusions can be made. The theoretical fits to the data of the two reactions indicate

a common value for λ of about 0.7 eV and widely differing optimal rates, about 10^9 s^{-1} and 10^2 s^{-1} , respectively. The $\hbar\omega$ for each reaction can be evaluated because the $-\Delta G^\circ$ dependences of rate has been determined at several temperatures between 295 K and 5 K (ref. 13). The examples of data in Fig. 3a are shown fitted to a gaussian curve with several values of $\hbar\omega$. The 35 K data demonstrate the requirement for low-temperature measurements in revealing a lower limit for $\hbar\omega$. Analysis of both data sets over the temperature range suggests a common $\hbar\omega$ of about 70 meV , which, at more than twice the 25 meV Boltzmann energy (kT) at ambient temperature, describes the significant involvement of protein high-frequency vibrational modes coupled to electron transfer¹³. Electric field effects on the Q_A^- to BChl_2^+ electron-transfer reaction rate confirm the involvement of high-frequency modes²⁰.

Figure 3b shows the $-\Delta G^\circ$ dependence of electron-transfer rates for several synthetic and semi-synthetic systems in solution at 295 K. For synthetic systems with donors and acceptors separated by $5\text{--}13 \text{ \AA}$ by different covalently attached bridging structures¹⁶⁻¹⁹, the fits to the data show λ values in the range $0.8\text{--}1.3 \text{ eV}$, with optimal rates between 5×10^9 and $5 \times 10^{11} \text{ s}^{-1}$. The data from all four bridged molecules at 295 K accommodate

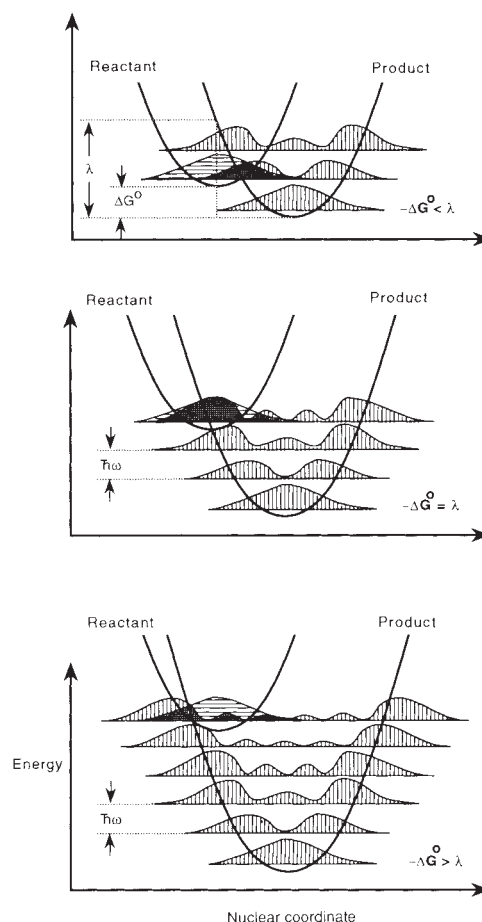


FIG. 1 Intersecting parabolic potential energy surfaces are used to represent classical reactant and product harmonic oscillators for three electron-transfer reactions with different free energies. Quantum mechanical harmonic oscillator wavefunctions are illustrated superimposed on these parabolas. Franck-Condon factors are proportional to the overlap of reactant nuclear wavefunctions (horizontal shading) and product wavefunctions (vertical shading). Overlap will be maximal when the free energy of the reaction ($-\Delta G^\circ$) matches the reorganization energy (λ). λ , Held constant in this figure, represents the energy required to move along the product potential surface as the nuclear geometry is distorted from the average product configuration to the average reactant configuration. Harmonic oscillator energy levels are separated by the quantum energy $\hbar\omega$. For clarity, only the lowest energy-reactant vibrational state is shown, corresponding to a Boltzmann thermal energy kT small compared with $\hbar\omega$.

BOX 1 Equations and parameters of electron transfer

Equation/symbol

Definition

$$k_{\text{et}} = \frac{2\pi}{\hbar} V_R^2 FC \quad (1)$$

Electron-transfer rate: Fermi's golden rule derived from quantum mechanical perturbation theory applied to non-adiabatic reactions

V_R^2	Electronic coupling	Quantum mechanical matrix element coupling reactant and product electronic wavefunctions.
FC	Franck-Condon-weighted density of states	Integrated overlap of reactant and product nuclear wavefunctions of equal energy.

$\hbar$ Planck's Constant

$$V_R^2 = V_0^2 \exp(-\beta R) \quad (2)$$

Electronic coupling reflecting simple exponential decay of electronic wavefunctions with distance

V_0^2	Maximum electronic coupling	
R	Distance	Centre of edge atom of donor to centre of edge atom of acceptor.
β	Beta	Exponential coefficient of decay of electronic coupling with R .

$$FC = (4\pi\lambda kT)^{-1/2} \exp[-(-\Delta G^\circ - \lambda)^2/4\lambda kT] \quad (3a)$$

Marcus classical form of overlap of harmonic oscillator wavefunctions with identical frequencies³

λ	Reorganization energy	Energy required to distort nuclear configuration of product state into the geometry of the reactant state without electron transfer. Generally rises with the increased polarity of the redox centre environment.
$-\Delta G^\circ$	Free energy	Difference between reactant and product state
k	Boltzman constant	

$$FC = (\hbar\omega)^{-1} \exp[-S(2n+1)] \left(\frac{n+1}{n}\right) \frac{P}{2} I_p[2S\sqrt{n(n+1)}] \quad (3b)$$

Quantum-corrected Marcus-type expression of Levich and Dogonadze⁷ and of Jortner⁸. Other expressions explicitly combine quantum high-frequency and classical low-frequency motion⁸.

$\hbar\omega$	Characteristic frequency	Reorganization energy-weighted average of frequencies of harmonic oscillators coupled to electron transfer, in energy units. Vibrations involving solvent, protein medium and redox cofactors range from <10 meV to >200 meV.
$S = \lambda/\hbar\omega$		Characteristic frequency normalized λ .
$P = -\Delta G^\circ/\hbar\omega$		Characteristic frequency normalized free energy.
$n = [\exp(\hbar\omega/kT) - 1]^{-1}$		Average harmonic oscillator vibrations level populated at temperature T .

$I_p()$ Modified Bessel function of order P

$$FC = (2\pi\sigma^2)^{-1/2} \exp[-(\Delta G - \lambda)^2/2\sigma^2] \quad (3c)$$

Semiclassical Hopfield expression⁹. This simple gaussian form is usually quite similar to equation (3b)

$\sigma^2 = \lambda \hbar\omega \coth(\hbar\omega/2kT)$	Variance of distribution with quantum correction for low temperatures.
---	--

values for $\hbar\omega$ of about 70 meV. A closer assessment of $\hbar\omega$ is possible for the triptycene-bridged molecule because of a recent kinetic determination in methyltetrahydrofuran glass at 77 K (open pentagons at top left of Fig. 3b)²¹. The freezing causes λ to diminish to ~0.6 eV but, as expected, the optimal rate remains similar. The breadth of the gaussian curve at 295 K and 77 K indicates that $\hbar\omega$ is comparable to that found in the reaction centre. The only other measurements of $-\Delta G^\circ$ dependency of rates at low temperature (77 K) have been on donors and acceptors dispersed in methyltetrahydrofuran glass²²; this also reveals an $\hbar\omega$ higher than ambient kT , in this case about 100 meV. Thus, reactions in which rates have been measured over an extensive $-\Delta G^\circ$ range suggest that the breadth of the roughly gaussian dependency on $-\Delta G^\circ$ curves at ambient temperatures is largely due to differences in the value of λ rather than to variations in $\hbar\omega$.

Figure 3b also shows data for the non-physiological reactions between the haem and ruthenated surface histidines of modified cytochromes *c*, each from a different species^{14,23}, and myoglobin¹⁵. Because the rates were obtained at $-\Delta G^\circ$ values approaching λ , the projected optimal rates are within a roughly 5–10-fold uncertainty.

It seems that the altered chemical nature of the donor and acceptor in these examples does not affect the Franck-Condon factors significantly. For biological systems, this conclusion is dramatically demonstrated by the replacement in the photo-synthetic reaction centre of the Q_A site quinone by the non-quinonoid compounds *m*-dinitrobenzene, 2-hydroxy-1-nitroso-naphthalene or fluorenones. The rates observed (Fig. 3a) for both reactions at 295 K and 35 K compare well with those obtained for quinones with the same $-\Delta G^\circ$ (K.W. and P.L.D., manuscript in preparation). Similarly, for the synthetic systems, several of the data sets in Fig. 3b, especially those from refs 16 and 22, are in fact composites of chemically different redox cofactors which evidently conform to the same Marcus-type expression. Furthermore, Fig. 3b presents excited and ground-state electron-transfer reactions^{14,17,18} within a single Marcus relationship, which suggests that V_0 , λ and $\hbar\omega$ may not be significantly different for singlet and triplet excited state or ground-state electron transfer.

Figure 4 shows excited and ground-state electron-transfer reactions for which the distance is established, but the rates are known only at limited $-\Delta G^\circ$ values. To compare these reactions with those with a more extensive $-\Delta G^\circ$ dependence, we fitted the available data to a Marcus-type relation using the 70 meV value for $\hbar\omega$ suggested above to predict a possible λ and optimal rate. Fits of electron-transfer rates for charge separation and recombination provide reasonable estimates for λ and the optimal rate in many systems. Two examples are shown as dashed curves in Fig. 4 for the covalently linked porphyrin-quinone incorporating zero (upper) and one (lower) spirocyclobutane unit²⁴. Analogous fits to electron-transfer rates from the singlet or triplet state of $\text{BChl}_2^+\text{BPh}^-$ in *Rb. sphaeroides* reaction centres to form the respective ground or triplet state BChl_2 (refs 25–29) suggest a λ of ~500 meV and an optimal rate of $2 \times 10^9 \text{ s}^{-1}$; similar values can be deduced for the same reaction in *Rp. viridis*²⁸.

The mode of electron transfer from BChl_2^* to BPh in *Rb. sphaeroides* is still controversial. One mechanism proposes BChl to be a conventional redox intermediate^{30,31}, whereas another uses BChl as a virtual intermediate in superexchange³². We confine our attention here to the simpler two-step model. Although the $-\Delta G^\circ$ values of the two individual steps are not known, the sum (160 meV) is small. Molecular dynamic simulations³³ suggest that the λ for the two steps is also small (~220 meV), which secures the native rates close to the optimal values, irrespective of the individual $-\Delta G^\circ$ values. This view is supported experimentally by the small dependence of the rates of these reactions on ΔG° modulation by applied electric fields (ref. 34, and C.C.M. *et al.*, manuscript in preparation). Although

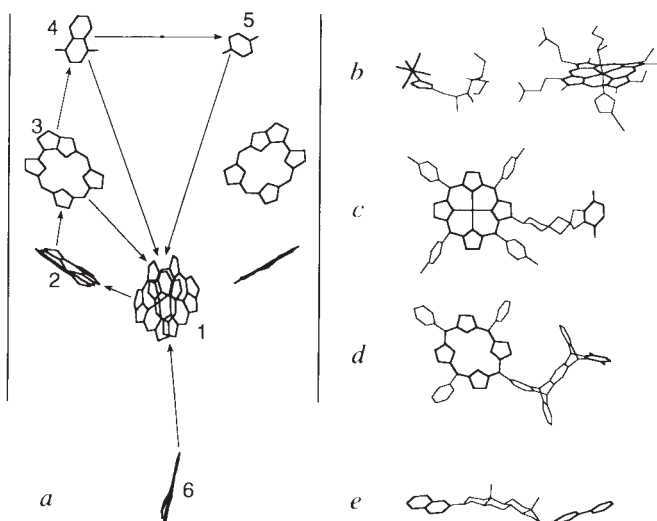


FIG. 2 Representative structures of electron-transfer systems with well defined donor-acceptor distances. *a*, The geometry of the cofactors of the photosynthetic reaction center of *Rb. viridis* revealed by X-ray diffraction¹⁰. After light excitation of the bacteriochlorophyll dimer (BChl₂; 1) electron transfer progresses sequentially through bacteriochlorophyll monomer (BChl; 2) bacteriopheophytin (BPh; 3), primary quinone (Q_A; 4), and secondary quinone (Q_B; 5). The resultant positively charged oxidized dimer (BChl₂⁺) can be rereduced by the nearest of four bound haem cytochrome *c* (6). Arrows indicate the physiologically productive charge separations and non-productive charge recombinations considered in this report. Vertical lines represent the rough dimension of the membrane profile across which the light-driven electron transfer generates transmembrane potential. The two cofactors to the right of the structure are BChl and BPh, symmetric to 2 and 3, do not seem to be involved as redox intermediates and their function is uncertain. *b*, Semisynthetic system of ruthenated histidine 39 (left) and zinc-substituted haem (right) of a cytochrome *c* from *Candida krusei*¹⁴. As with the reaction centre structure, most of the amino acid-groups have been removed for clarity. *c*, Porphyrin-quinone pair linked covalently by a saturated spirocyclobutane bridge containing one cyclobutane unit²⁴. *d*, Porphyrin-quinone pair linked covalently by the pentiptycene bridge containing four aromatic rings⁴⁷. *e*, Naphthalene-biphenyl pair linked covalently by a steroid bridge and activated by pulse radiolysis¹⁶. Donor and acceptor redox centres are illustrated with thicker lines than the amino acids and bridging groups.

not shown, very similar plots are obtained using data from *Rp. viridis*³⁵.

Figure 4 presents three other photosynthetic reactions, the rates of which are known only at relatively small $-\Delta G^\circ$ and have a correspondingly larger uncertainty in the extrapolated optimal rates. These are electron transfer from Q_A⁻ to Q_B (ref. 36), from Q_B⁻ back to BChl₂⁺ in a mutant of *Rb. sphaeroides* chosen to obviate interference from rate-limiting proton exchange³⁷, and from cytochrome *c*₅₅₉ to BChl₂⁺ (ref. 38). The points fitted to the Marcus-type relation with the 0.7–1.3 eV range of values for λ , and 70 meV value for $\hbar\omega$ provided by the reactions discussed, bracket estimates for the optimal rates. The reactions, their distances and optimal rates are summarized in Table 1.

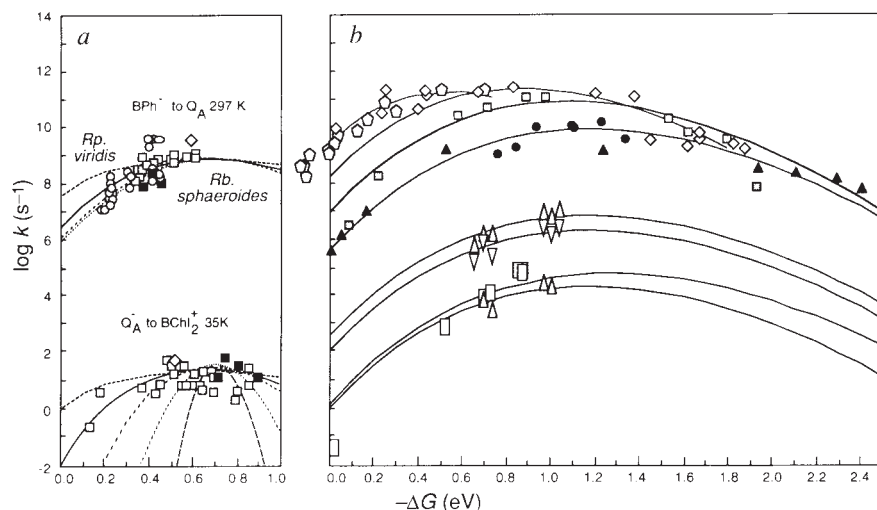
Electronic terms

Figure 5*a* describes the logarithm of the optimal rates compiled from the variety of electron-transfer reactions in protein as a function of the edge-to-edge distances between donor and acceptor (Table 1). The optimal rates defined by the reactions with extensive free-energy dependence describe a remarkable linearity that is further supported by the other photosynthetic reactions. This linearity shows that the donor and acceptor electronic wavefunctions not only decay exponentially but do

so similarly through distinct intervening protein media of two reaction centres and several haem-proteins. The relationship yields an electron-transfer rate of 10^{13} s^{-1} at van der Waals contact. This short-range electron-transfer rate corresponds to the characteristic frequency of the nuclei considered in the Franck-Condon term. The slope of the line provides a value of β of about $1.4 \text{ \AA}^{-1}$, equivalent to a 10-fold change in rate for a $1.7 \text{ \AA}$ change of distance. By comparison, elementary tunnelling calculations suggest that the electron-transfer rate through vacuum ($\beta \sim 2.8 \text{ \AA}^{-1}$)³⁹ would fall off about 10 times faster than through protein for each $1.7 \text{ \AA}$ of distance.

Figure 5*b* compares the optimal rate-distance relationship for electron transfer in a variety of covalently bridged donor-acceptor molecules (Table 1) with that of protein (solid line). There is substantial overlap of rates at distances between 5 and $10 \text{ \AA}$. At longer distances the optimal rates in the covalent molecules are faster, whereas at van der Waals contact the projected rate is slower. The trend of the data (dashed line) suggests an average β of about $0.7 \text{ \AA}^{-1}$, equivalent to a 10-fold change in rate for a $3.3 \text{ \AA}$ change of distance. As with protein systems, the synthetic systems present an order of magnitude spread in optimal rates compared at similar distances. This spread makes it difficult to determine whether aromatic linkers themselves accelerate electron-transfer rates and introduces a

FIG. 3 Free-energy dependence of electron-transfer rates in various systems. *a*, BPh⁻ to Q_A and Q_A⁻ to BChl₂⁺ in photosynthetic reaction centres. The free energy was changed by replacing the native Q_A with other compounds with different electrochemistry. *Rb. sphaeroides*: native quinone (◇), non-native quinones (□)¹³, non-quinonoids (■) (K.W. *et al.*, manuscript in preparation); *Rp. viridis*: native quinone (○), non-native quinones (○) (J.M.K. *et al.*, manuscript in preparation). The data are fit using the quantized Marcus expression given by equation (3b) in Table 1 with $\lambda = 0.7 \text{ eV}$ for both reactions and increasingly sharper curves with $\hbar\omega = 150 \text{ meV}$ (dotted), 70 meV (solid), 30 meV (large-dashed), and 10 meV (small-dashed). Equation (3a) the classical Marcus expression, is shown for comparison as the narrowest curve in the 35 K Q_A⁻ to BChl₂⁺ data (dashed). The corresponding expression for the 297 K BPh⁻ to Q_A data is similar to the 10 meV curve and is not included. *b*, Semisynthetic ruthenium-cytochrome *c* (▽)²³, (Δ)¹⁴ and ruthenium-myoglobin systems (□)¹⁵; bridged porphyrin-quinone (◇)¹⁸, (Δ)²¹, (●)¹⁹; biphenyl-acceptor (▲)¹⁶ and diiridium-pyridinium (⊗)¹⁷ molecules. A common value of $\hbar\omega$ of 70 meV with the



published values of λ were used to fit all the data. Experiments were done near room temperature unless specified.

TABLE 1 Optimal rates and distances of Fig. 5

Description Reaction centres (Fig. 5a, circles)	Species	Distance (Å)	log 10 rate (s ⁻¹)	Reference
BChl ⁻ to BPh	<i>Rb. sphaeroides</i>	4.6	12.2	31
BChl ⁻ to BPh	<i>Rp. viridis</i>	4.8	12.2	35
BChl ₂ ⁺ to BChl	<i>Rb. sphaeroides</i>	5.4	11.5	31
BChl ₂ ⁺ to BChl	<i>Rp. viridis</i>	5.5	11.5	35
BPh ⁻ to BChl ₂ ⁺	<i>Rp. viridis</i>	9.5	9.6	28
BPh ⁻ to Q _A	<i>Rp. viridis</i>	9.6	9.0	*
BPh ⁻ to Q _A	<i>Rb. sphaeroides</i>	10.0	9.0	13
BPh ⁻ to BChl ₂ ⁺	<i>Rb. sphaeroides</i>	10.1	9.6	26, 27
cyt c ₅₅₉ to BChl ₂ ⁺	<i>Rp. viridis</i>	12.3	8.2	38, 43
Q _A ⁻ to Q _B	<i>Rp. viridis</i>	13.5	7.6	29, 44
Q _A ⁻ to Q _B	<i>Rb. sphaeroides</i>	14.5	6.6	36
Q _A ⁻ to BChl ₂ ⁺	<i>Rp. viridis</i>	22.4	1.8	45
Q _A ⁻ to BChl ₂ ⁺	<i>Rb. sphaeroides</i>	22.5	1.2	13
Q _B ⁻ to BChl ₂ ⁺	<i>Rb. sphaeroides</i>	23.4	0.2	37
Semisynthetic (Fig. 5a, triangles)	Protein			
Zn haem/His 33 Ru	Cyt c	12.4	6.7	23
Zn haem/His 39 Ru	Cyt c	13.0	7.1	14
Zn haem/His 62 Ru	Cyt c	15.5	4.6	14
Haem/His 81 Ru	Myoglobin	19.3	3.0	15
Haem/His 116 Ru	Myoglobin	20.1	2.9	15
Haem/His 12 Ru	Myoglobin	22.0	2.9	15
Synthetic (Fig. 5b)	Major bridging group			
Porphyrin/quinone	Spirocyclobutane (0)	5.0	11.2	24
Ir ₂ /pyridinium	Alkylphosphinite	5.8	11.2	17
Porphyrin/quinone	Triptycene	5.8	11.5	18
Porphyrin/quinone	Phenyl-cyclopentane	6.0	10.8	46
Porphyrin/quinone	Penttiptycene	7.1	10.2	47
Porphyrin/quinone	Penttiptycene	7.3	10.7	47
Porphyrin/quinone	Phenyl-adamantane	8.0	9.6	46
Porphyrin/quinone	Spirocyclobutane (1)	8.2	10.3	24
Re/dimethylaniline	Peptide	8.4	10.1	48
Porphyrin/quinone	Penttiptycene	8.6	10.4	47
Porphyrin/quinone	Penttiptycene	9.1	10.9	47
Ir ₂ /pyridinium	Arylphosphinite	9.4	10.2	†
Porphyrin/quinone	Spirocyclobutane (2)	9.6	9.6	24
Porphyrin/quinone	Phenyl-bicyclooctane	10.1	9.9	19
Porphyrin/quinone	Norbornyl-bicyclohexane	11.1	9.2	49
Biphenyl/acceptor	Steroid	11.7	9.7	16
Porphyrin/quinone	Spirocyclobutane (3)	12.6	8.5	24
Re/thiafulvene	Cyclohexane	13.6	9.1	50

* J.M.K. *et al.*, and † R.S.F. *et al.*, manuscripts in preparation.

note of caution to the interpretation of site-directed mutagenic substitution of aromatic and non-aromatic protein residues. The simple electron-transfer model does not address the source of the spread in rates or differences in the projected rate at van der Waals contact.

The β value of 1.4 Å^{-1} for protein, intermediate between that found for covalent systems and vacuum, can be understood in terms of the degree of structural continuity between the donor and acceptors. In this sense the protein matrix seems to resemble a frozen organic glass, such as methyltetrahydrofuran, in which the β value has been measured as 1.2 Å^{-1} (ref. 22). A theoretical approach on the basis of estimates of wavefunction propagation factors for covalent bonds, hydrogen bonds and through-space jumps along specific paths in proteins generates β values less and greater than 1.4 Å^{-1} (ref. 40). But Fig. 5a suggests that the practical effects on rate from specific gaps and links in the protein medium are small, perhaps in part because the relatively large size of most biological redox centres offers a number of competing paths and leads to a fairly uniform β .

Biological engineering considerations

The design and engineering of redox proteins draws on the nuclear and electronic terms presented. Although the values of $\hbar\omega$ and β are important in defining the nature of electron transfer

in redox proteins, the observation in Fig. 5a that the entire range of physiologically productive and unproductive electron transfer in the photosynthetic reaction centre, and non-physiological electron transfer in distinct haem proteins, can be accommodated by a single $\hbar\omega$ and β , suggests that these terms are basic properties of protein and do not modulate electron-transfer rates to any physiological significance. The necessarily combined effects of $-\Delta G^\circ$ and λ can in principle modulate the rate up to 10^5 -fold over the extreme range of 0–1.3 eV shown in Figs 3 and 4. In practice, the modulation of rate by $-\Delta G^\circ$ for any λ is significantly constrained by the demands of most biological reactions to conserve energy, together with the shallowness of the dependence of the rate on $-\Delta G^\circ$. By comparison, distance is a freely variable parameter that provides a 10^{12} -fold range of rates in the examples presented and seems to be primarily responsible for the promotion of desired, physiologically productive electron transfer and the suppression of competing non-productive events in most biological electron-transfer reactions.

The partnership of distance, $-\Delta G^\circ$ and λ can be seen at work in elementary electron-transfer reactions involving cytochrome *c* and *c*₂. The distance between redox centres revealed by X-ray and modelled structures for the complexes of cytochrome *c* and cytochrome *c* peroxidase⁴¹ and of cytochrome *c*₂ and BChl₂

(ref. 42) has apparently been selected to be 12–16 Å to accommodate the prevailing $-\Delta G^\circ$ and large λ values⁴¹ while maintaining sub-millisecond rates sufficient to compete effectively with the lifetime of the encounter complexes.

Similar considerations apply to successful electron transfer across the 35 Å span of biological membrane in respiratory complexes such as ubiquinol-cytochrome *c* oxidoreductase or the cytochrome *c* and ubiquinol oxidases. For an electron to cross the entire span using two cofactors, say, from the centre of one haem to another, the edge-to-edge distance will be between 25 and 35 Å depending on orientation of the haems, pushing the optimal electron-transfer rate into the seconds and days timescale, respectively. The $>10^3 \text{ s}^{-1}$ overall rates of respiratory systems can only be achieved if at least one other intervening redox cofactor is involved, irrespective of $-\Delta G^\circ$ and λ values. Indeed, all the characterized transmembrane redox proteins use at least three cofactors, including two porphyrins. Functional studies suggest that the distances between cofactors positioned in the membrane profile are small enough to effect transmembrane electron transfer in less than milliseconds using $-\Delta G^\circ$ values in the zero to 250 meV range. Time constraints for electron transfer in reaction centres are more severe because of the roughly 10^9 s^{-1} decay rate of the initial excited state. Thus, any photosystem requires more closely spaced redox centres with at least three steps and four redox centres for efficient transmembrane electron transfer; the reaction centre of Fig. 2a uses five redox centres.

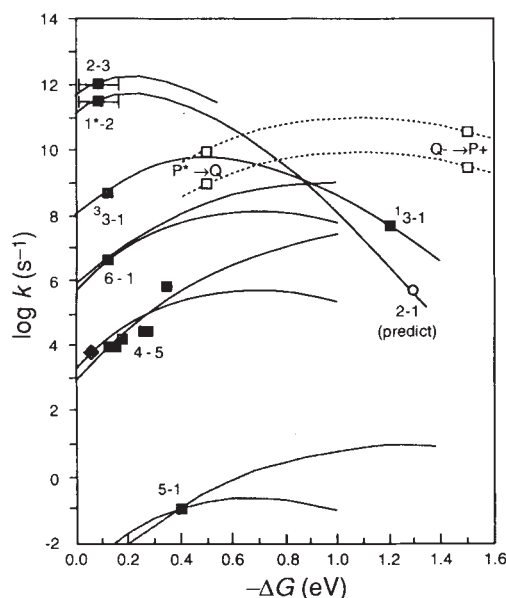


FIG. 4 Free-energy-rate relations for electron-transfer reactions in the photosynthetic reaction centre for which the experimental free-energy dependence is limited. The curves are derived using the Marcus-type expression (equation (3b)) from Box 1 with the $\hbar\omega$ value of 70 meV obtained in Fig. 3 and a range of values of λ . The electron-transfer reactions are numbered as in Fig. 2a. The reactions presented include: charge recombination of the singlet and triplet $\text{BPh}^-\text{BChl}_2^+$ in *Rb. sphaeroides* (1-2 and 3-1) with a common λ of 0.5 eV; cytochrome c_{559} to BChl_2^+ in *Rp. viridis* (6-1) with a λ range of 0.7 to 1.0 eV; Q_A^- to Q_B in *Rb. sphaeroides* for a variety of quinone substitutions (4-5) and Q_B^- to BChl_2^+ for a pH-insensitive mutant of *Rb. sphaeroides* (5-1) with possible values of λ ranging from 0.7–1.3 eV. A two-step initial charge separation mechanism is considered; both steps BChl_2^+ to BChl (1-2) and BChl^- to BPh (2-3) are shown with a 220 meV reorganization energy and include the uncertainty of the individual $-\Delta G^\circ$ in the overall 160 meV. An anticipated rate for $\text{BChl}^- \text{BChl}_2^+$ ground-state charge recombination (2-1) is included in the 'inverted region' of the BChl_2^+ to BChl curve (open circle). Two examples of reactions involving synthetic bridged porphyrin-quinone charge separation ($\text{P}^* \rightarrow \text{Q}$) and recombination ($\text{Q}^- \rightarrow \text{P}$) rates are shown for comparison; the open squares and dashed lines describe rates for the spirocyclobutane bridge containing one (lower) and no (upper) cyclobutane unit. References are included in Table 1.

When distances are similar, Franck-Condon factors can have a critical role in defining the direction of electron transfer in biological systems. This is especially clear in photosynthetic systems that face a special engineering problem not present in respiratory systems: initial charge separation proceeds at nearly 100% quantum efficiency, even though charge recombination back to the ground state presents a similar distance and a large favourable $-\Delta G^\circ$. Although several plausible explanations for directional specificity involving the special characteristics of the excited state and charge-separated electronic wavefunctions have been advanced³², we suggest that such specificity can also be accommodated simply by Franck-Condon factors alone. The initial charge separation seems designed to take place with small λ and $-\Delta G^\circ$ values compared with the overall light energy of 1.4 eV, pushing the large $-\Delta G^\circ$ of the direct back-reaction into the Marcus 'inverted region'. As predicted on Fig. 4 (reaction 2-1), this has the effect of retarding direct charge recombination and assuring efficient, productive electron transfer. The inverted region may also have a role in assuring efficient electron transfer from BPh^- to Q_A by slowing the competing charge recombination from BPh^- to BChl_2^+ , despite the similar distance and hence similar optimal rate. Franck-Condon factors could be critical

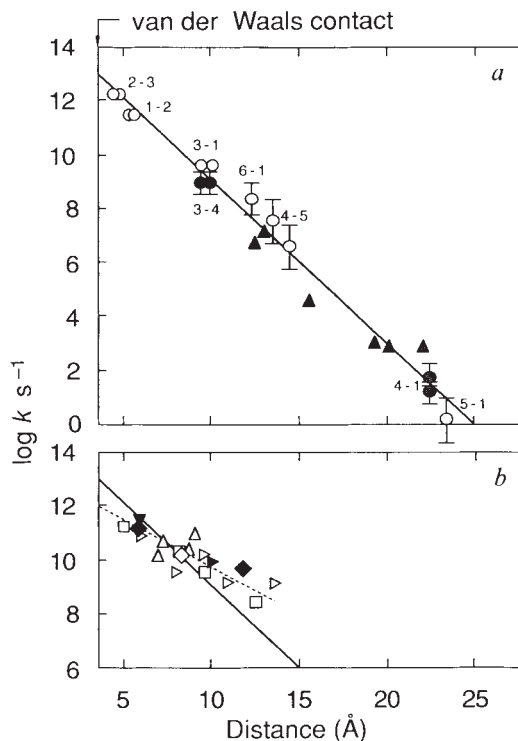


FIG. 5 The free-energy optimized rate versus edge-to-edge distance relationship for electron transfer. a, Protein, including the photosynthetic reaction centre (○) and semi-synthetic Ru-cyt *c* and Ru-Mb (△); filled symbols are for reactions presented in Fig. 3. The numbers refer to the cofactors involved in photosynthetic charge separation in the reaction centre: BChl_2 (1), BChl (2), BPh (3), Q_A (4), Q_B (5) and cyt *c* (6). Error bars are associated with uncertainties in rate optima: for cyt *c* to BChl_2^+ the error bars span a λ range of 0.7 to 1.0 eV; for Q_A^- to Q_B and Q_B^- to BChl_2^+ which may have quite polar environments, the error bars span a λ range of 0.7 to 1.3 eV. b, Covalently linked systems. Triangles represent bridges incorporating aromatic groups: triptycene (▽), pentyptycene (△) and others (◇). Quadrilaterals represent bridges without aromatic groups: spirocyclobutanes (□) and others (◇). Data from Fig. 3 are shown as filled symbols. The fit line of a represents a simple exponential decay of equation (2) with β of 1.4 Å^{-1} . Distance is defined as center of edge atom of donor to center of edge atom of acceptor; thus the vertical line at 3.6 Å represents van der Waals contact. For reaction centres these distances are derived from the crystal structure^{10–12}. Distances in covalent systems were measured with LabVision software and reflect typical shortest edge-to-edge values among energy-minimized rotational isomers. Descriptions of individual points are given in Table 1.

in guiding electron transfer through one branch of the nearly symmetric arrangement of chlorins visible in the reaction centre structure of Fig. 2a. The observed high quantum efficiency of electron transfer through the active branch would be assured if formation of BChl_2^+ BChl^- in the inactive branch has a λ of about 600 meV or, as already suggested³³, an endothermic ΔG° of 200 meV.

Although we have emphasized the importance of engineering electron transfer within a maximal distance set by time constraints, it is equally important to provide a minimum distance to the protein exterior to insulate against unwanted reactions. For example, the H subunit capping the reaction

centre seems to perform this role by providing 20 Å insulation between the Q_A and Q_B radicals and any adventitious redox centres.

We believe that natural selection has shaped the present form of electron-transfer proteins by applying the engineering principles we have outlined, building on the characteristic $\hbar\omega$ and β of the biological medium and adjusting the three parameters ΔG° , λ and especially distance. This biological 'blueprint' sketches the central features of the function of some of the most important electron-transfer systems in nature and clarifies the design requirements for the construction of analogous systems in the laboratory. □

Received 27 November 1991; accepted 21 January 1992.

- Chance, B. & Nishimura, M. *Proc. natn. Acad. Sci. U.S.A.* **46**, 19–24 (1960).
- Devault, D. & Chance, B. *Biophys. J.* **6**, 825–847 (1966).
- Mitchell, P. *Nature* **191**, 144–148 (1961).
- Devault, D. *Q. Rev. Biophys.* **13**, 387–564 (1980).
- Marcus, R. A. *J. chem. Phys.* **24**, 966–978 (1956).
- Marcus, R. A. & Sutin, N. *Biochim. biophys. Acta* **811**, 265–322 (1985).
- Levich, V. G., & Dogonadze, R. R. *Dokl. Akad. Nauk USSR* **124**, 123–126 (1959).
- Jortner, J. *J. chem. Phys.* **64**, 4860–4867 (1976).
- Hopfield, J. J. *Proc. natn. Acad. Sci. U.S.A.* **71**, 3640–3644 (1974).
- Michel, H., Deisenhofer, J. & Epp, O. *EMBO J.* **5**, 2445–2451 (1986).
- Allen, J. P., Feher, G., Yeates, T. O., Korniya, H. & Rees, D. C. *Proc. natn. Acad. Sci. U.S.A.* **84**, 5730–5734 (1987).
- Chang, C. H. *et al.* *FEBS Lett.* **205**, 82–86 (1986).
- Gunner, M. R. & Dutton, P. L. *J. Am. chem. Soc.* **111**, 3400–3412 (1989).
- Therien, M. J. *et al.* in *Electron Transfer in Inorganic, Organic and Biological Systems* (eds Bolton, J., McLendon, G. L. & Mataga, N.) 191–199 (American Chemical Society, Washington DC 1991).
- Cowan, J. A., Upmacis, R. K., Beratan, D. N., Onuchic, J. N. & Gray, H. B. *Ann. N. Y. Acad. Sci.* **550**, 68–84 (1988).
- Closs, G. L. & Miller, J. R. *Science* **240**, 440–447 (1988).
- Fox, L. S., Kozik, M., Winkler, J. R. & Gray, H. B. *Science* **247**, 1069–1071 (1990).
- Wasielowski, M. R., Niemczyk, M. P., Svec, W. A. & Pewitt, E. B. *J. Am. chem. Soc.* **107**, 1080–1082 (1985).
- Joran, A. D. *et al.* *Nature* **327**, 508–511 (1987).
- Franzen, S., Goldstein, R. F. & Boxer, S. G. *J. phys. Chem.* **94**, 5135–5149 (1990).
- Gaines, G. L., O'Neil, M. P., Svec, W. A., Niemczyk, M. P. & Wasielowski, M. R. *J. Am. chem. Soc.* **113**, 719–721 (1991).
- Miller, J. R., Beitz, J. V. & Huddleston, R. K. *J. Am. chem. Soc.* **106**, 5057–5068 (1984).
- Meade, T. J., Gray, H. B. & Winkler, J. R. *J. Am. chem. Soc.* **111**, 4353–4356 (1989).
- Knapp, S. *et al.* *J. Am. chem. Soc.* **113**, 4010–4013 (1991).
- Parson, W. W., Clayton, R. K. & Codgell, R. J. *Biochim. biophys. Acta* **387**, 265–277 (1975).

- Schenck, C. C., Blankenship, R. E. & Parson, W. W. *Biochim. biophys. Acta* **680**, 44–59 (1982).
- Chidsey, C. E. D., Kirmaier, C., Holten, D. & Boxer, S. G. *Biochim. biophys. Acta* **766**, 424–437 (1985).
- Holten, D., Windsor, M. W., Parson, W. W. & Thornber, J. P. *Biochim. biophys. Acta* **501**, 112–126 (1978).
- Shopes, R. J. & Wraight, C. A. *Biochim. biophys. Acta* **806**, 348–356 (1985).
- Shuvalov, V. A. & Klevanik, A. V. *FEBS Lett.* **160**, 51–55 (1983).
- Holzapfel, W. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **87**, 5168–5172 (1990).
- Michel-Beyerle, M. E. *et al.* *Biochim. biophys. Acta* **932**, 52–70 (1988).
- Parson, W. W., Chu, Z. T. & Warshel, A. *Biochim. biophys. Acta* **1017**, 251–272 (1990).
- Lockhart, D. J., Kirmaier, C., Holten, D. & Boxer, S. G. *J. phys. Chem.* **94**, 6987–6995 (1990).
- Holzapfel, W. *et al.* *Chem. Phys. Lett.* **160**, 1–7 (1989).
- Giangiacomo, K. M. & Dutton, P. L. *Proc. natn. Acad. Sci. U.S.A.* **86**, 2658–2662 (1989).
- Takahashi, E. & Wraight, C. A. *FEBS Lett.* **283**, 140–144 (1991).
- Shopes, R. J., Holten, D., Levine, L. & Wraight, C. A. *Photosyn. Res.* **12**, 165–180 (1987).
- Gamow, G. Z. *Phys.* **51**, 204–212 (1928).
- Beratan, D. N., Betts, J. N. & Onuchic, J. N. *Science* **252**, 1285–1288 (1991).
- McLendon, G. *Accs chem. Res.* **21**, 160–167 (1988).
- Tiede, D. M. *et al.* in *The Photosynthetic Bacterial Reaction Center, Structure and Dynamics* (eds Breton, J. & Vermeglio, A.) 13–20 (Plenum, New York, 1988).
- Dracheva, S. M. *et al.* *Eur. J. Biochem.* **171**, 253–264 (1988).
- Carithers, R. P. & Parson, W. W. *Biochim. biophys. Acta* **387**, 194–211 (1975).
- Shopes, R. J. & Wraight, C. A. *Biochim. biophys. Acta* **893**, 409–425 (1987).
- Wasielowski, M. R. & Niemczyk, M. P. in *Porphyrins: Excited States and Dynamics* (eds Gouterman, M., Rentzepis, P. M. & Straub, K. D.) 154–165 (Am. Chem. Soc., Little Rock, Arkansas, 1985).
- Wasielowski, M. R., Niemczyk, M. P., Johnson, D. G., Svec, W. A. & Minsek, D. W. *Tetrahedron* **45**, 4785–4806 (1989).
- MacQueen, D. B. & Schanze, K. S. *J. Am. chem. Soc.* **113**, 7470–7479 (1991).
- Antolovich, M. *et al.* *J. phys. Chem.* **95**, 1933–1941 (1991).
- Perkins, T. A., Hauser, B. T., Eyler, J. R. & Schanze, K. S. *J. phys. Chem.* **94**, 8745–8748 (1990).

ACKNOWLEDGEMENTS. This work was supported by grants from the NSF and the NIH.

LETTERS TO NATURE

Optical counterpart of the east radio lobe of M87

M. Stiavelli[†]*, J. Biretta[‡], P. Møller^{*} & W. W. Zeilinger^{*}

* European Southern Observatory, Karl-Schwarzschildstrasse 2, D-8046 Garching bei München, Germany

† Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56100 Pisa, Italy

‡ National Radio Astronomical Observatory, PO Box 0, Socorro, New Mexico 87801, USA

RADIO galaxies are divided into two broad classes¹ according to whether they show extended radio lobes ending in hotspots (compact bright sources of radio emission) at some distance from the nucleus (Fanaroff–Riley class II) or instead have bright regions or knots close to the nucleus (FR I). Hotspots and knots emit synchrotron radiation, sometimes even in the optical range^{2,3}. The galaxy M87 (NGC4486) in Virgo is the closest giant elliptical galaxy showing a jet and double-lobe radio structure, and is classified as FR I mainly because of the morphology of the jet-dominated west radio lobe, even though the east radio lobe resembles typical FR II objects and shows a hotspot⁴. We report here the detection of the optical counterpart of this hotspot. Because the lifetime of optically emitting synchrotron electrons is short, we suggest that this hotspot is fed by an unseen counterjet. This conclusion adds weight to the idea that apparently one-sided jet sources are intrinsically two-sided, with the visibility of the receding jet reduced by beaming effects.

We have directly imaged the central regions of M87 at high resolution at the 2.5-m Nordic Optical Telescope (NOT) and at the 2.2-m European Southern Observatory/Max Planck Institute telescope. The observations at the NOT were done on 18 May 1991 with the Stockholm CCD camera. Exposures in the V (5 min), R (5 min) and I (10 min) bands were secured under conditions of exceptional seeing (full width at half maximum, FWHM < 0.6"). The ESO/MPI 2.2-m images were taken with the ESO Faint Object Camera and spectrograph 2 (EFOSC-2) on 8 June 1991 in the U (30 min), B (15 min) and V (1.5 min) filters under conditions of mediocre seeing (FWHM ≈ 1.3"). Details of the observations and data reduction are described elsewhere (ref. 5, and W.W.Z., P.M. and M.S., manuscript in preparation). An important part of the data reduction is the subtraction of a smooth model for the stellar light of M87. This model has been obtained by fitting elliptical isophotes to each image. The overall accuracy of this subtraction was better than 1%.

Our optical maps cover all the area where bright radio emission is observed⁶, and show an arclike feature roughly opposite to the jet at a distance of ~24 arcsec from the nucleus (see Fig. 1a). The emission is conspicuous in all bands with the exception of the U filter, where the detection is only at the 2 σ level. This emission 'arc' is ~4.5 arcsec long in the north–south direction, but is unresolved in the east–west direction in our best image, which has 0.58-arcsec FWHM seeing (see Fig. 1b). We can exclude the possibility that this feature is an H α filament, as it is visible in the B, V and I images. In addition, no prominent H α filament is visible at that location in narrow-band images⁷,

TABLE 1 Fluxes

Filter	Log ν (Hz)	Magnitude	Flux (mJy)
U	14.9	21.8 $\pm$ 0.4	0.003 $\pm$ 0.0012
B	14.8	21.7 $\pm$ 0.2	0.009 $\pm$ 0.002
V	14.7	21.0 $\pm$ 0.2	0.015 $\pm$ 0.003
R	14.6	20.3 $\pm$ 0.2	0.024 $\pm$ 0.004
I	14.5	19.4 $\pm$ 0.1	0.036 $\pm$ 0.004
2 cm	10.17	—	238 $\pm$ 7

and the known, prominent H α filaments are much fainter than this source in our broad-band images.

Comparison with low-resolution radio images⁴ shows that the optical feature is coincident in position with the radio emission maximum in the east radio lobe (Fig. 1a, feature θ). We have also compared the optical image with a higher resolution (0.15" FWHM) 15-GHz radio image (J.B., manuscript in preparation), which affords greater contrast between feature θ and the diffuse radio emission. This shows that the morphologies of the optical emission and the brightest radio emission (>1.0 mJy beam⁻¹) are essentially identical.

We have determined the spectral index of the arc by integrating the flux in our maps with a circular aperture with 4 arcsec diameter. The measured fluxes are given in Table 1. We obtain an optical spectral index ($S_\nu \propto \nu^{-\alpha}$) $\alpha = 1.9 \pm 1.0$ (excluding the U filter data), consistent with the one measured at the outer edge of the jet⁸ (knots H and H'). For the purpose of measuring the radio-to-optical spectral index we integrated the radio flux density in the same aperture on a 15-GHz map. This map was observed (J.B., in preparation) with the Very Large Array in the A, B and C configurations, and was convolved to 0.6-arcsec FWHM resolution. In an effort to better isolate the radio emission from the arc, background levels were determined 5 arcsec to the east and west, averaged and then subtracted from the radio image. The resulting 15-GHz flux density was 238 ± 7 mJy, which gives a radio-to-optical spectral index of 0.88 ± 0.01 . These observed spectral indices resemble those measured in the hot-spots of FR II galaxies. For instance, 3C33 South⁹ has an optical spectral index of 1.97 and a radio-to-optical index of 0.94. This consideration, and the morphological similarity with the radio synchrotron emission, suggest that the optical continuum emission is simply an extension of the radio synchrotron spectrum.

The minimum-pressure analysis carried out by Hines *et al.*⁴ can be repeated for the larger emissivity implied by this extension of the synchrotron spectrum. By adopting radio spectral indices in the range 0.6–0.8 and assuming a spectral break in the range 10^{13} – 10^{14} Hz, we derive a total power of $\sim 8 \times 10^{40}$ erg s⁻¹ (with an emissivity density ~ 7 times as high as in ref. 4), a magnetic field at the lobe of 88 μ G and a corresponding lifetime for synchrotron losses (for electrons emitting in the B band) of $\sim 1,600$ yr. Because of this short lifetime, it seems unlikely that the feature can be attributed to a transient phenomenon. Furthermore, transient effects such as shocks or turbulence seem unlikely to produce the high emissivity contrast seen in the θ without some form of continual driving, such as from a jet⁴.

Models in which the electrons are either supplied or locally reaccelerated by an 'unseen' jet then seem necessary. (This is in contrast to the radio-emitting electrons which have lifetimes of $\sim 10^6$ years, and can be attributed to events in the distant past or flow from other regions of the source.) The maximum distance that can be covered by electrons emitted from the nucleus is ~ 5 kpc, when the Compton losses caused by the M87 radiation field (energy density $\sim 10^{-11}$ erg cm⁻³) and the synchrotron losses on the background field⁴ (22 μ G) are considered. As the projected distance of feature θ from the core is only 1.7 kpc, a direct replenishment of relativistic electrons from the nucleus cannot be excluded. The energy lost to synchrotron emission by these electrons during their journey from the core

must, however, be smaller than $\sim 5\%$ of the energy emitted in θ . The limit on the ratio of jet to counterjet brightness⁵ (not including θ) is in fact ~ 450 , and θ is only ~ 20 times fainter than the jet itself (100 times at optical wavelengths). By comparison, the energy 'lost' in the western jet before reaching knot A is larger than 10% of that emitted there. Alternatively, the optically emitting electrons could be supplied by particle acceleration within θ , if it is associated with shocks at the end of an unseen jet.

The apparent absence of a counterjet might be attributed either to beaming effects, or to intrinsic differences between the visible jet and counterjet. If θ is powered by a relativistic counterjet directed away from the observer, Doppler beaming could render the counterjet undetectable. The observed arc might then represent a region where the jet fluid has been slowed to subrelativistic speeds through interaction with the surrounding medium. If the jet and oppositely directed counterjet are intrinsically similar in brightness, such a model requires the jet to be within $\sim 30^\circ$ of the line of sight and have a fluid velocity $>0.8c$ (ref. 10). These parameters do seem plausible, and large proper motions ($>0.6c$) have been observed in the jet (ref. 11).

The areas immediately to the north and south of θ show radio emission at 40% of the one measured in θ itself with 4" apertures, and in the optical the measured flux south of θ is 0.006 ± 0.004 in the I band instead of the expected 0.014. Although only at the 2σ level, this supports the idea that this area does not have the same spectral index as θ . Unfortunately no optical flux can

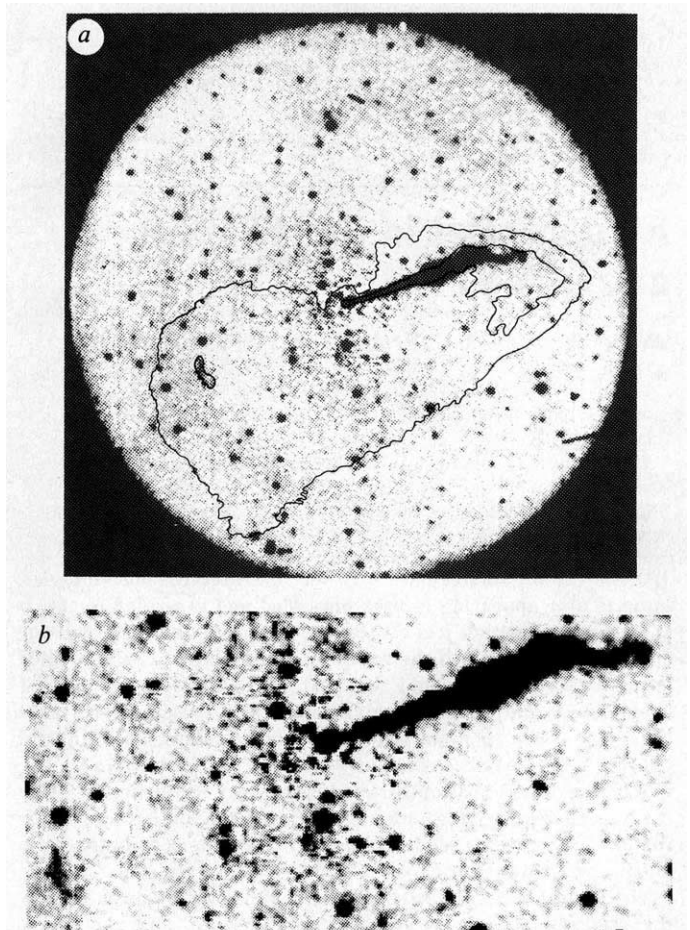


FIG. 1 a, Combined frame obtained by averaging all the NOT V, R and I images after subtraction of a smooth model for the galaxy light profile. The arclike feature identified as the optical counterpart of the east radiolobe feature θ of ref. 4 appears opposite to the jet, perpendicular to the jet axis. Contours⁴ taken by the Very Large Array at radio wavelength 6 cm at intensities of 1 and 6 mJy per beam are shown for reference. b, A higher-contrast central section shows the arclike shape of the detected emission.

be derived north of θ , because of the presence of a star. If optical emission were confirmed over these regions north and south of the prominent arc, it would suggest there are at least some intrinsic differences between the jet and counterjet, such as the opening angle.

In any case, the hotspot in M87 is not too different from what is observed in some 'jetless' hotspots in FR II objects (for example in PKS0521-36). Its optical identification supports the idea that this galaxy is an undersized and underpowered classical double-lobe radio galaxy. Its apparent peculiarity may be caused either by the small intrinsic power of the central engine or by the interaction with a dense gaseous environment. The sharp and localized emission from hotspot θ , extending to optical wavelengths, its alignment with the nucleus and the western jet, and its morphological similarity with a shock front perpendicular to the jet axis, suggest that this feature is still powered by an 'invisible' jet or by what remains of a now-extinct eastern jet. By understanding how this counterjet can transfer large amounts of relativistic electrons out to a few kiloparsecs without being seen, we would probably be able to understand how the same can happen on the much larger spatial scales of giant FR II radiogalaxies. $\square$

Received 3 October; accepted 12 December 1991.

1. Fanaroff, B. L. & Riley, J. M. *Mon. Not. R. astr. Soc.* **167**, 31-35p (1974).
2. Tyson, J. A., Crane, P. & Saslaw, W. C. *Astr. Astrophys.* **59**, L15 (1977).
3. Crane, P., Tyson, J. A. & Saslaw, W. C. *Astrophys. J.* **265**, 681-700 (1983).
4. Hines, D. C., Owen, F. N. & Eilek, J. A. *Astrophys. J.* **347**, 713-726 (1989).
5. Stivalelli, M., Møller, P. & Ziegler, W. W. *Nature* **354**, 132-134 (1991).
6. Owen, F. N., Hardee, P. E. & Bignell, R. C. *Astrophys. J.* **239**, L11-15 (1980).
7. Ford, H. C. & Butcher, H. *Astrophys. J. Suppl.* **41**, 147 (1979).
8. Biretta, J. A., Stern, C. P. & Harris, D. E. *Astr. J.* **101**, 1632-1646 (1991).
9. Crane, P. & Stockton, A. in *Proc. Workshop on Hot Spots in Extragalactic Radio Sources* (eds Meisenheimer, K. & Röser, H.-J.) 115-117 (1988).
10. Biretta, J., Owen, F. & Cornwell, T. *Astrophys. J.* **342**, 128-134 (1989).
11. Biretta, J. & Owen, F. in *Parsec Scale Radio Jets* (eds Zensus, J. A. & Pearson, T. J.) (Cambridge University Press, 1990).

A counterjet in the elliptical galaxy M87

William B. Sparks, D. Fraix-Burnet*, F. Macchetto & F. N. Owen†

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA

* Observatoire de Toulouse, 14 Avenue Édouard Belin, F-31400, Toulouse, France

† National Radio Astronomy Observatory, PO Box 0, Socorro, New Mexico 87801, USA

JETS in radio galaxies have a variety of observed morphologies, in particular appearing in both one-sided and two-sided forms. It is not known whether jets can be intrinsically one-sided, or if the one-sided appearance is the result of an asymmetry in a two-sided jet. An asymmetry could be real, or apparent because of geometrical and possibly relativistic effects creating a large difference in the apparent visibility of the two sides. The elliptical galaxy M87 hosts a weak but well known one-sided radio jet, which is very visible at optical and even X-ray wavelengths. Using the New Technology Telescope of the European Southern Observatory in Chile, we have identified an optical counterpart to the radio hotspot in the southeast lobe of M87, at about the same distance as the radio jet but oppositely located. Polarization data identify the optical emission as synchrotron radiation, and the severe energetic requirements of maintaining this emission suggest continuous replenishment at the site of the hotspot. We propose that this is provided by an invisible 'counterjet'. We argue that the counterjet cannot be identical to the visible known jet, but must be intrinsically somewhat fainter.

The radio jet in M87 progresses smoothly out in a cone to Knot A, after which oscillations occur and the apparent jet

direction eventually reverses as it disperses into the NW radio lobe^{2,3}. Even with radio data of very high dynamic range (of order $1:10^5$) and optical CCD images, there is no sign of a similar jet on the opposite side^{4,5}. Plausible explanations for this asymmetry are that there is no counterjet (intrinsic asymmetry), or that there is a counterjet but we do not see it because it is highly beamed away from us (apparent asymmetry) or because radiative losses along the jet are small (intrinsic asymmetry).

We obtained optical images of M87 with the New Technology Telescope of the European Southern Observatory (ESO) at La Silla, Chile, during April 1990. Exposures of 5 min at V and R were obtained with the ESO Faint Object Camera and Spectrograph 2 (EFOSC-2) which used a thinned back-illuminated RCA CCD chip whose pixel size on the sky was measured to be ~ 0.26 arcsec. Seeing was ~ 0.7 arcsec full width at half maximum for R and 0.8 arcsec for V. In addition, imaging polarimetric observations had been acquired during February 1990 with the ESO 3.6-m telescope with EFOSC-1, set up as described by Sparks *et al.*⁶. These images comprised direct integrations of 10 min at B, 2.5 min at V and four exposures each of 5 min using the V filter in combination with optical polarizers mounted at 45° position angles within the collimator of EFOSC-1. The seeing for the B observation was 1.8 arcsec, whereas for the observation of V with polarizers it was 1.3 arcsec.

To study structure other than the galaxy itself, we fitted ellipses to the isophotes of the galaxy and constructed an empirical model of the underlying stellar light which had exactly the isophotal profile of the galaxy. An iterative rejection criterion in the ellipse fitting caused jet influence on derived ellipses to be minimal. The models were subtracted from the imaging data. The resulting images were linearly resampled onto a finer pixel grid, with pixels of side 0.16 arcsec and blinked on an image display device against the Very Large Array image of ref. 3, which was also resampled onto the same pixel grid. The correspondence between the radio and optical features at the location designated θ by Hines *et al.* was striking (Figs 1 and 2). The feature is 24 arcsec from the nucleus of M87 at position angle 113° . It consists of a bright spot in the northern corner of a triangle of emission and has a total extent in the optical of $\sim 3 \times 5$ arcsec. The bright peaks in the radio and optical images are exactly coincident as far as can be determined from the data.

To obtain a flux calibration, we scaled the galaxy intensity profiles to match the B, V and R profiles of Young *et al.*⁷. For the peak of the hotspot, counts were integrated in the model-subtracted images within a square aperture of side 1.44 arcsec

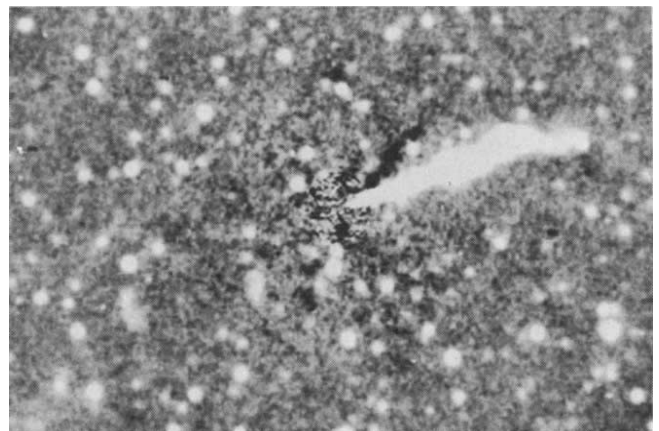


FIG. 1 NTT image of M87, $67'' \times 45''$, after subtraction of a model of the underlying galaxy. The optical feature in question is at a radius of 24 arcsec from nucleus (base of the jet) at a position angle 113° . The diffuse object closer to the nucleus is probably a background galaxy. The dust absorption north of the jet will be discussed elsewhere (W.B.S., H. C. Ford and A. Kinney, manuscript in preparation).

TABLE 1 Flux estimates and spectral indices for hotspot

Wavelength	Exposure (s)	Seeing (arcsec)	Peak mag	Flux (Log Jy)	α^*	Total mag	Flux (Log Jy)	α^*
6-cm VLA				-1.09	0.63†		-0.37	0.71†
R (NTT)	300	0.7	21.51	-5.15		20.00	-4.55	
V (NTT)	300	0.8	22.14	-5.27	0.83‡	20.66	-4.68	0.86‡
V (3.6-m)	630§	1.4	22.17	-5.28		20.91	-4.78	
B (3.6-m)	600	1.8	22.54¶	-5.39		21.63	-5.02	

* Spectral index is α where $S_\nu \propto \nu^{-\alpha}$ from 6 cm to optical.

† Estimated from flux at 20 cm, 6 cm and 2 cm.

‡ Estimated from 6 cm to V.

§ Effective exposure: sum of four 5-min V plus polaroid, plus direct 2.5-min V.

|| Seeing correction 0.55 mag included.

¶ Seeing correction 0.89 mag included.

centred at the position of the maximum radio surface brightness. The total flux from the optically emitting region was estimated by integrating counts within an area 4.0×5.0 arcsec. To allow for slight errors in the local background introduced by subtraction of the model, sky regions on either side of the object were used to adjust the derived counts appropriately.

Table 1 presents magnitudes for the bright peak and total emission patch separately. At radio wavelengths, the peak has a flatter spectral index, α , than the patch as a whole, 0.63 compared with 0.71, although these are uncertain as it is necessary to subtract a contribution from diffuse-lobe radio emission. The spectral indices from 6 cm to V are 0.83 and 0.86, whereas the optical data alone indicate steeper indices of $\alpha \approx 1.2$ for the peak and $\alpha \approx 2.3 \pm 0.5$ for the total emission. For the diffuse radio emission outside the optical patch, the lower limit to the spectral index from 6 cm to V was $\alpha \geq 0.95$, determined in two ways: by interactively scaling and subtracting the 6-cm and V-band images, and by measuring the r.m.s. variations in the V sky data. Anything with shallower spectrum should have been visible, suggesting that the diffuse emission has a steeper radio-to-optical spectrum than the hotspot.

Given the probability of optical synchrotron emission, it was appropriate to investigate the polarization. We assumed that the galaxy starlight is unpolarized, and we derived linear relations between the intensities in each polarization image by plotting the median galaxy intensities in circular annuli between 5 and 40 arcsec sampled at one pixel intervals, against one another. The polarization of the peak, measured in this way, was $\sim 34 \pm 5\%$ at a position angle of the magnetic vector $\phi \approx 30 \pm 5^\circ$, which is $\sim 83^\circ$ from the radius vector. The polarization of the whole region, in contrast, was insignificant in the optical. Formally, the polarization percentage was $P \leq 5\%$. One of the images contains a small spike within the hotspot which may be due to a cosmic ray. Replacing the value of that pixel by the mean of the surrounding pixels gives a revised polarization of 30% at position angle $\phi \approx 27^\circ$, which is 86° from the radius vector.

The light travel time from the nucleus along the jet as seen in projection is $\sim 10^4$ yr. The synchrotron lifetime of optically emitting electrons, given by $\sim 1.8 \times 10^4 B^{-3/2} \nu^{-1/2}$ yr, where B is the magnetic field strength in gauss and ν is the frequency in hertz, is only a few hundred years or even less in the M87 jet². This means that continuous acceleration of electrons at the

optically emitting site is almost certainly required. Given that the emission is almost exactly opposite the jet ($182 \pm 1^\circ$) the most likely source of such an energy supply is a spatially symmetric counterjet, rather than, for example, a chance interaction between the interstellar medium and the radio lobe. This is supported by the increasing spectral index with distance from the peak, with an energy injection site identified as the peak of the hotspot and synchrotron ageing effects as one moves away from it. The optical feature shows essentially all of the characteristics of a hotspot⁹, conventionally interpreted as the 'working surface' at the end of a jet. The jet seems to end at this knot *θ*, whereas on the opposite side (corresponding to knot *H*) the jet continues outward. The radio-to-optical spectral index is typical of knots in jets¹⁰ and the W hotspot in Pic A has roughly the same spectral index¹¹. The position angle of polarization is indicative of a magnetic field orthogonal to the line to the nucleus, also consistent with other hotspots in radio lobes. The low polarization for the patch as a whole may arise if there is unresolved polarization structure within it, with polarization vectors at different orientations cancelling one another out.

The question remains as to why we do not see the SE portion of the jet. If the jet is intrinsically one-sided and has recently flipped from the SE to the NW side, then because we still see optical emission, it can have done so only in the very recent past (a few hundred years ago). Otherwise, the optical emission would have decayed. But as the whole length of the NW side of the jet is visible, with a light travel time of $\sim 10^4$ yr, this seems implausible.

Is relativistic beaming sufficient to diminish its intensity below detectability? Laing¹² and Garrington *et al.*¹³⁻¹⁵ observed in a sample of one-sided jets that higher Faraday rotation in the lobe occurs on the side without the jet. They suggest that this effect is due to a longer path length through a galactic halo for the side without the jet and therefore that it is at a greater distance from us, hence favouring the beaming interpretation of one-sidedness. In the case of M87, there are two problems with this. First, polarization studies of the radio source¹⁶ point to a Faraday screen enveloping each lobe rather than the whole source and second, other arguments^{2,4,5} show that the apparent jet motion is only mildly relativistic, $\beta < 0.5$, and that the jet lies close to the plane of the sky, $< 45^\circ$. An intrinsic asymmetry in the radiative losses of the jet therefore seems to be the most likely option.

One possibility is that a dissipationless channel has been established on the SE side, unlike on the NW side. This would seem odd given that radio polarization observations indicate a higher Faraday depth on the SE side. But the detailed morphology of the NW jet does indicate a hollow, emission-free channel along the centre of the jet², and an extension of the physical processes that enable dissipationless energy flow along that route may provide the solution.

A second possibility is that the synchrotron emission process in the jet differs from one side to the other, depending on the interaction between the jet and the interstellar medium¹⁷. This idea is supported by a triple correlation between the depolarization of the lobes, the spectral index of the lobes and the visibility of the jets, indicating intrinsically asymmetric properties. The strict applicability to the case of M87 is beyond our scope here,

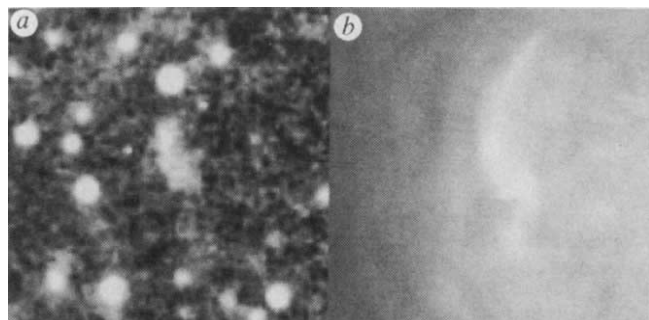


FIG. 2 *a*, Section from Fig. 1. *b*, Same region at the same scale extracted from the VLA image of Hines *et al.*³, for comparison.

but a more detailed model will be presented elsewhere (D.F.B., manuscript in preparation). The observations therefore suggest the existence of a spatially symmetric SE counterjet, similar in scale to the NW optical jet, but with an intrinsic asymmetry in its visibility. $\square$

Received 11 October; accepted 26 December 1991.

1. Bridle, A. H. in *Physics of Energy Transport in Extragalactic Radio Sources Proc. NRAO Workshop No. 9* (eds Bridle, A. H. & Eilek, J. A.) (National Radio Astronomy Observatory, 1984).
2. Owen, F. N., Hardee, P. E. & Cornwell, T. J. *Astrophys. J.* **340**, 698–707 (1989).
3. Hines, D. C., Owen, F. N. & Eilek, J. A. *Astrophys. J.* **347**, 713–726 (1989).
4. Reid, M. J., Biretta, J. A., Junor, W., Muxlow, T. W. B. & Spencer, R. E. *Astrophys. J.* **336**, 112–120 (1989).
5. Biretta, J. A., Owen, F. N. & Cornwell, T. J. *Astrophys. J.* **342**, 128–134 (1990).
6. Sparks, W. B., Paresce, F. & Macchetto, F. *Astrophys. J.* **347**, L65–68 (1989).
7. Young, P. J., Westphal, J. A., Kristian, J., Wilson, C. P. & Landauer, F. P. *Astrophys. J.* **221**, 721–730 (1978).
8. Rybicki, G. B. & Lightman, A. P. *Radiative Processes in Astrophysics* (Wiley, New York 1979).
9. Laing, R. in *Hot Spots in Extragalactic Radio Sources* (eds Meisenheimer, K. & Röser, H.-J.).
10. Fraix-Burnet, D. et al. *Astr. J.* **101**, 88–93 (1991).
11. Meisenheimer, K. et al. *Astr. Astrophys.* **219**, 63–86 (1989).
12. Laing, R. A. *Nature* **331**, 149–151 (1988).
13. Garrington, S. T., Leahy, J. P., Conway, R. G. & Laing, R. A. *Nature* **331**, 147–149 (1988).
14. Garrington, S. T., Conway, R. G. & Leahy, J. P. *Mon. Not. R. astr. Soc.* **250**, 171–197 (1991).
15. Garrington, S. T. & Conway, R. G. *Mon. Not. R. astr. Soc.* **250**, 198–208 (1991).
16. Owen, F. N., Eilek, J. A. & Keel, W. C. *Astrophys. J.* **362**, 449–454 (1990).
17. Fraix-Burnet, D. in *Proc. 7th IAP Astrophysics Meeting* (eds Pelletier, G., Roland, J. & Sol, H.) (Cambridge University Press, 1991).

Generation of ultrasonic waves from a layered photoacoustic source

T. Sun & G. J. Diebold

Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA

ACOUSTIC pulses are used extensively for imaging, materials characterization and non-destructive testing. Here we describe a method for producing trains of square-wave photoacoustic pulses. Our device comprises an alternating series of light-absorbing and transparent fluid layers. When this layered structure is irradiated with an amplitude-modulated laser, constructive and destructive interference of the resulting photoacoustic waves generates pressure waves at specific resonance frequencies, the bandwidths of which are determined by the number of layers and their acoustic impedances. When operated at the resonance condition, a device consisting of a large number of layers produces ultrasonic waves that are highly directional in space. A structure excited by a single short pulse from a Nd:YAG laser generates a train of acoustic square waves, the pulsewidth of each being determined by the transit of sound across the absorbing layer.

The process of absorption of optical radiation by a body and the subsequent thermal expansion and emission of an acoustic wave is known as the photoacoustic effect¹. Irradiation of an optically thin fluid layer by a short laser pulse has recently been shown to generate a series of square waves in time². If a photoacoustic source consisting of optically transparent layers of one fluid interposed between source layers of a second fluid is constructed so that the transit time of sound across each layer is identical, the sound waves reflected at the interfaces of the layers can add in phase with those from the source layers, resulting in a coherent generation of sound. The response of such a device to amplitude-modulated optical radiation is determined by solution of the wave equation for pressure^{3,4}, which, together with the acoustic conservation equations⁵ applied at the interfaces of the layers, shows that the n th layer produces left- and right-going plane waves at the ends of the device with amplitudes A_n and B_n , respectively, the values of which are determined as solutions to the matrix equation (T.S. and G.J.D.,

manuscript in preparation; see also refs 6, 7)

$$[\alpha\beta]^{N-n}[\alpha_0]\begin{bmatrix} 0 \\ B_n \end{bmatrix} - \begin{bmatrix} \kappa_f/\hat{q} \\ \kappa_f/\hat{q} \end{bmatrix} = [\xi]\left\{[\alpha^*\beta^*]^{n-1}[\alpha_0^*]\begin{bmatrix} A_n \\ 0 \end{bmatrix} - \begin{bmatrix} \kappa_f/\hat{q} \\ \kappa_f/\hat{q} \end{bmatrix}\right\} \quad (1)$$

where $\hat{q}$ is a dimensionless modulation frequency, $\hat{q} = 2\pi fl/c$ where f is the modulation frequency, l is the thickness of the absorbing layer and c is the sound speed in the absorbing layer. Note that phase factors have been absorbed into A_n and B_n here. The matrices α and α_0 propagate waves from the absorbing medium to the transparent medium, and the matrix β propagates the wave from the transparent medium to the absorbing medium. Both matrices are a function of the acoustic impedance ratio for the two media (the density ratio $\hat{\rho}$ times the sound speed ratio $\hat{c}$) and their thicknesses. Here, ξ is a diagonal matrix which is a function of the absorbing medium's wavevector and thickness, and κ_f is a source amplitude dependent on the laser intensity and thermal properties of the absorbing layers. To determine the total wave amplitude as a function of frequency, equation 1 must be solved for each source at a series of discrete frequencies over the desired frequency range, and the solutions for each of the sources added.

The acoustic pressure from a structure with five absorbing layers plotted as a function of the dimensionless modulation frequency $\hat{q}$ displays a series of resonances (Fig. 1a) located at frequencies corresponding to constructive interference in the wave amplitudes from each of the N layers. When the number of layers is increased (see Fig. 1b), both the widths of the peaks and the amplitudes of the side lobes decrease as a result of the increase in the number of reflections at the interfaces of the two media. A similar narrowing is seen when the parameter $\hat{\rho}\hat{c}$ is increased, an effect which can be viewed as a reduction in the coupling efficiency of acoustic energy between the layers and the surrounding medium.

Experimental photoacoustic signals were generated by a structure formed of sheets of coloured polymethylmethacrylate

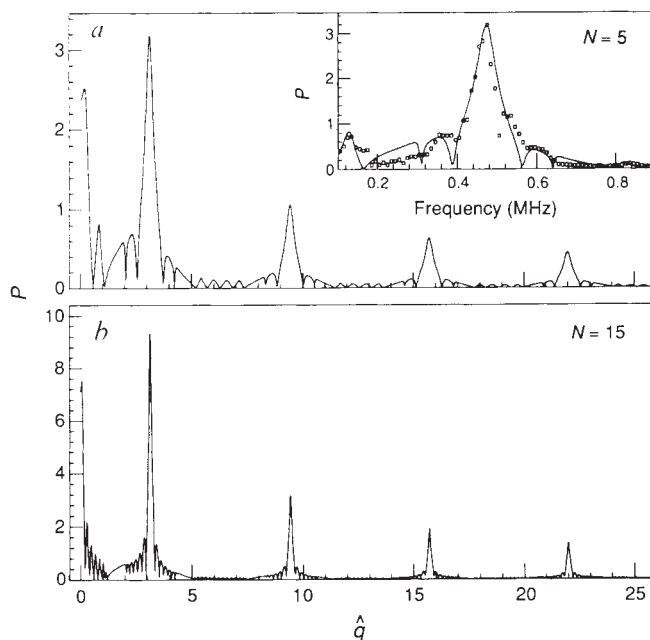


FIG. 1 Photoacoustic pressure P (in arbitrary units) against dimensionless frequency parameter $\hat{q}$ calculated from equation (1) for (a) five absorbing layers and (b) 15 absorbing layers, with $\hat{\rho}\hat{c} = 2.1$. Inset: experimental photoacoustic pressure P (in arbitrary units) against modulation frequency f for five 2.75-mm-thick absorbing Lucite layers separated by 1.53-mm layers of distilled water. The parameter $\hat{\rho}\hat{c}$ in the calculated curve is 2.1, the value for absorbing Lucite and transparent water layers. The time constant on the lock-in amplifier was 10 s.

(Lucite) separated by distilled water. (An isotropic solid responds as a fluid provided the optically excited area is sufficiently large; M. I. Khan, T.S. and G.J.D., manuscript in preparation.) The layers were irradiated with 514.5-nm radiation from an argon laser, sinusoidally modulated by a Pockels cell and then passed through a beam expander and iris to give a 1-W beam, 3.5 cm in diameter. The photoacoustic waves were detected in water by a fast polyvinylidene film transducer^{8,9} connected to a preamplifier whose output was fed to a high-frequency lock-in amplifier. The inset in Fig. 1 shows the magnitude of the photoacoustic signal plotted against modulation frequency in the region of the first resonance. Higher-order resonances were found in other experiments.

The response of the structure to a short pulse of radiation can be found by numerical Fourier transformation of the solutions to equation (1). As shown in Fig. 2, the photoacoustic wave generated by irradiation of a layered structure with $\hat{\rho}\hat{c} = 9.4$ (corresponding to absorbing glass and transparent water layers) is an alternating series of compressions and rarefactions in time. As the number of layers or the parameter $\hat{\rho}\hat{c}$ is increased, the number of square-wave cycles becomes larger. Experimental waveforms (see Fig. 2), were generated by irradiating disks of coloured glass separated by water with the 8-ns output of a frequency-doubled Nd:YAG laser. The signal from the polyvinylidene film transducer was amplified by a factor of 40 (with a 5-MHz bandwidth preamplifier) and displayed on an oscilloscope.

The angular distribution of the acoustic radiation from n absorbing layers (juxtaposed with acoustically identical, transparent layers) excited by a laser beam of finite radius a can be found by evaluation of the integral solution to the wave equation, which gives the pressure as

$$P = \frac{ik_f \hat{a}^2}{2\hat{r}} \hat{q} \operatorname{sinc}\left(\frac{\hat{q} \cos \theta}{2}\right) \left[\frac{J_1(\hat{q} \hat{a} \sin \theta)}{\hat{q} \hat{a} \sin \theta} \right] \sum_{n=-\infty}^{\infty} e^{i\hat{q}(2n \cos \theta - \hat{r}_i)} \quad (2)$$

where $\hat{a} = a/l$, $\hat{r} = r/l$ (where r is the distance from the origin),

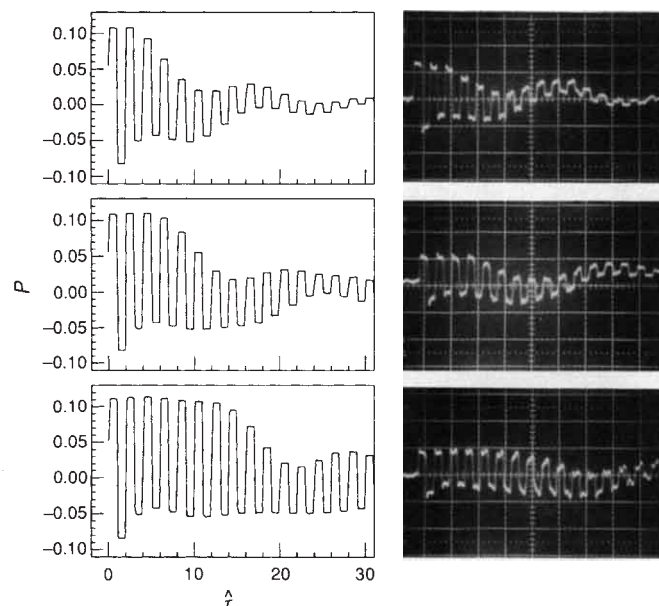


FIG. 2 Left column: photoacoustic pressure P (in arbitrary units) against dimensionless retarded time $\hat{\tau}$ from the edge of the absorbing layer furthest to the right, for (a) two layers, (b) three layers and (c) five layers with $\hat{\rho}\hat{c} = 9.4$. Right column: experimental oscilloscope traces of P against time from 3.26-mm-thick absorbing glass layers corresponding to the calculated waveforms in the left column for two, three and five absorbing layers. The time and voltage scales in the three experimental traces are 2 μ s per division and 20 mV per division. The 3.26-mm-thick glass layers were separated by 0.87-mm-thick water layers. The layers were irradiated with a laser beam of 150 mJ per pulse, expanded to a diameter of 8 cm.

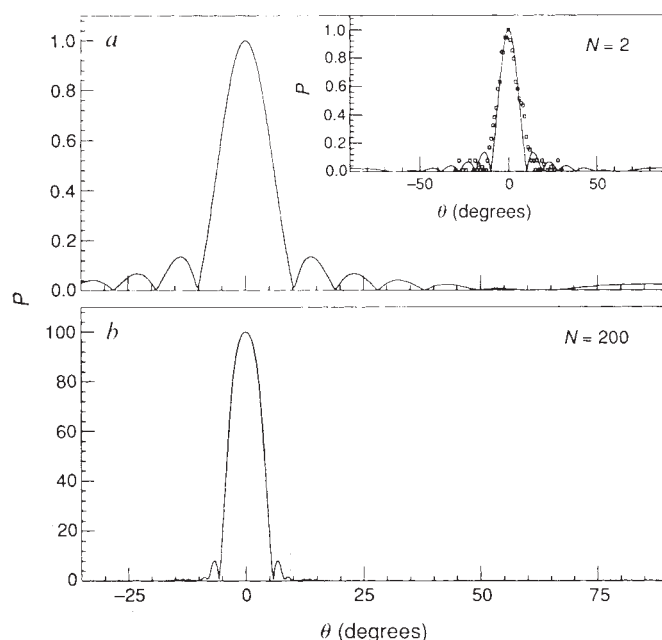


FIG. 3 Acoustic pressure P (in arbitrary units) against angle from the normal to the layers, θ , for $\hat{q} = \pi$ and $\hat{a} = 6.9$, calculated from equation (2) for (a) two layers and (b) 200 layers. Inset: experimental photoacoustic pressure against angle θ for two 1.6-mm-thick, absorbing water layers.

$\hat{r}_i$ is the dimensionless retarded time from the origin, J_1 is a Bessel function and θ is the angle measured from the normal to the layers. The curves in Fig. 3 show that increasing the number of absorbing layers increases the directionality of the acoustic beam; further calculations show that an arbitrarily high directionality can be obtained by making n sufficiently large. The inset in Fig. 3 shows an experimentally determined angular distribution of photoacoustic radiation at 460 kHz obtained from alternating layers of dyed and transparent water (separated by plastic film) irradiated by an expanded argon ion laser beam.

The frequency-domain photoacoustic response shown in Fig. 1 can be found in the optical transmission spectrum of a multi-layer dielectric interference filter¹⁰, or in the pressure response of an acoustic filter¹¹. In the first case alternation of the refractive indices of the layers plays the part of the differing acoustic impedances in the photoacoustic source. Of further note is the photoacoustic analogue of Fraunhofer diffraction¹⁰ in classical optics: the radiation pattern described by equation (2) in the limit of a single, thin absorbing layer reduces exactly to the Fraunhofer electric field distribution for a circular aperture. The summed exponential factor in equation (2) also implicitly contains the Bragg condition for X-ray diffraction by crystals¹² or optical diffraction by a grating¹⁰. Of course, the characteristic of the photoacoustic source that distinguishes it from its analogues is that it is an active device.

From the results given here it is possible to visualize a layered acoustic source excited by a train of optical pulses tailored so that the optical and acoustic pulse repetition rates are identical. Owing to the ease with which short optical pulses can be produced in a mode-locked laser, extension of the frequency range of the photoacoustic source to produce a highly collimated train of acoustic waves in the high megahertz range seems straightforward. □

Received 4 November 1991; accepted 7 January 1992.

1. Tam, A. C. *Rev. mod. Phys.* **58**, 381-431 (1986).
2. Diebold, G. J., Khan, M. I. & Park, S. M. *Science* **250**, 101-104 (1990).
3. Westervelt, P. J. & Larson, R. S. *J. acoust. Soc. Am.* **54**, 121-122 (1973).
4. Temkin, S. *Elements of Acoustics* (Wiley, New York, 1981).
5. Diebold, G. J. & Westervelt, P. J. *J. acoust. Soc. Am.* **84**, 2245-2251 (1988).
6. Burmistrova, L. V., Karabutov, A. A., Portnyagin, A. I., Rudenko, O. V. & Cherepetskaya, E. B. *Akust. Zh.* **24**, 655-663; (Engl. transl.) *Sov. Phys. Acoust.* **24**, 369-374 (1978).

7. B. Gusev, *Vestn. Moskov. Univ.* **22**, 7-12 (1981).
8. Tam, A. C. & Coufal, H. G. *Appl. Phys. Lett.* **42**, 33-35 (1983).
9. Coufal, H. *IEEE Trans. Ultrason. Ferro. Freq. UFFC* **33**, 507-512 (1986).
10. Born, M. & Wolf, E. *Principles of Optics* (Pergamon, New York, 1980).
11. Lindsay, R. B. *Mechanical Radiation* (McGraw-Hill, New York, 1960).
12. Schwartz, L. H. & Cohen, J. B. *Diffraction from Materials* (Springer, New York, 1987).

ACKNOWLEDGEMENTS. We are grateful for the support of the US Department of Energy, Office of Basic Energy Studies and thank the Alcoa Foundation for a research award.

A three-variable model of deterministic chaos in the Belousov-Zhabotinsky reaction

László Györgyi*† & Richard J. Field*

*Department of Chemistry, University of Montana Missoula, Montana 59812, USA

†Institute of Inorganic and Analytical Chemistry Eötvös Lóránd University, Budapest-112, PO Box 32 H-1518, Hungary

CHAOS is exhibited by a wide variety of systems governed by nonlinear dynamic laws¹⁻³. Its most striking feature is an apparent randomness which seems to contradict its deterministic origin. The best-studied chaotic chemical system is the Belousov-Zhabotinsky (BZ) reaction⁴⁻⁶ in a continuous-flow stirred-tank reactor (CSTR). Here we present a simple mechanism for the BZ reaction which allows us to develop a description in terms of a set of differential equations containing only three variables, the minimum number required to generate chaos in a continuous (non-iterative) dynamical system². In common with experiments, our model shows aperiodicity and transitions between periodicity and chaos near bifurcations between oscillatory and steady-state behaviour, which occur at both low and high CSTR flow rates. While remaining closely related to a real chaotic chemical system, our model is

sufficiently simple to allow detailed mathematical analysis. It also reproduces many other features of the BZ reaction better than does the simple Oregonator⁷ (which cannot produce chaos).

The temporal behaviour of a homogeneous chemical system is described by a set of often nonlinear ordinary differential equations. Oscillations and chaos may occur^{2,3} if appropriate feedback loops exist⁸ and if the system is maintained sufficiently far from thermodynamic equilibrium in a CSTR. The dynamics of the CSTR flow may couple with the chemical dynamics to increase the complexity of observed behaviour, as occurs in our model. The Belousov-Zhabotinsky⁶ (BZ) reaction is the Ce(IV)/Ce(III)-catalysed oxidation and bromination of CH₂(COOH)₂ by BrO₃⁻ in ~1M H₂SO₄. Its dynamic structure is well understood, and excellent chemical models of it are available^{9,10}.

The complex periodicity and chaos observed in BZ-CSTR experiments is characterized by the interaction of two frequencies¹¹ occurring on different timescales. We have developed an 11-variable model¹⁰ in which complexity arises because the cycling concentration of bromomalonic acid, a product of the natural BZ limit-cycle oscillations, acts as a bifurcation parameter between dominance of the BZ limit cycle and a stable steady state. The concentration of bromomalonic acid [BrMA], oscillates between values for which either the limit cycle or the steady state is stable. The basic BZ limit cycle defines the first frequency and the cycling BrMA concentration defines the second. The oscillations in concentration occur as small-amplitude ripples on top of a relatively high BrMA and result from the coupling of its oscillatory rate of production with its removal by the constant CSTR flow. The [BrMA] cycle does not become entrained with the BZ limit cycle because its frequency is defined by the CSTR flow rate as well as by the BZ chemistry. Its concentration is a bifurcation parameter because BrMA is a precursor of Br⁻, the control intermediate in the main feedback loop of the BZ reaction. Bifurcations between steady state and oscillatory behaviour occur at both high and low rates of Br⁻ production, and thus at both high and low [BrMA]. At high flow rates (low [BrMA]), an oxidized, high-[Ce(IV)]/[Ce(III)] and high-[HBrO₂] steady state appears, and at low flow rates (high [BrMA]), a reduced, low-[Ce(IV)]/[Ce(III)] and low-[HBrO₂] steady state appears. Simple oscillatory behaviour occurs at intermediate Br⁻ production rates. Complex oscillations and chaos occur near the bifurcations. These complex oscillations are usually a mixture of small- and large-amplitude oscillations, which we describe by the notation n^k , where n is the number of large-amplitude oscillations, and k is the number of small-amplitude oscillations per cycle. A cycle can consist of several units of small- and large-amplitude oscillations, leading to notations such as $1^4 1^5 1^5$.

The four-variable chemical mechanism shown in Table 1 is a reduction of the 11-variable model¹⁰ and is the same as one derived previously¹² except that we have deleted a Br⁻ produc-

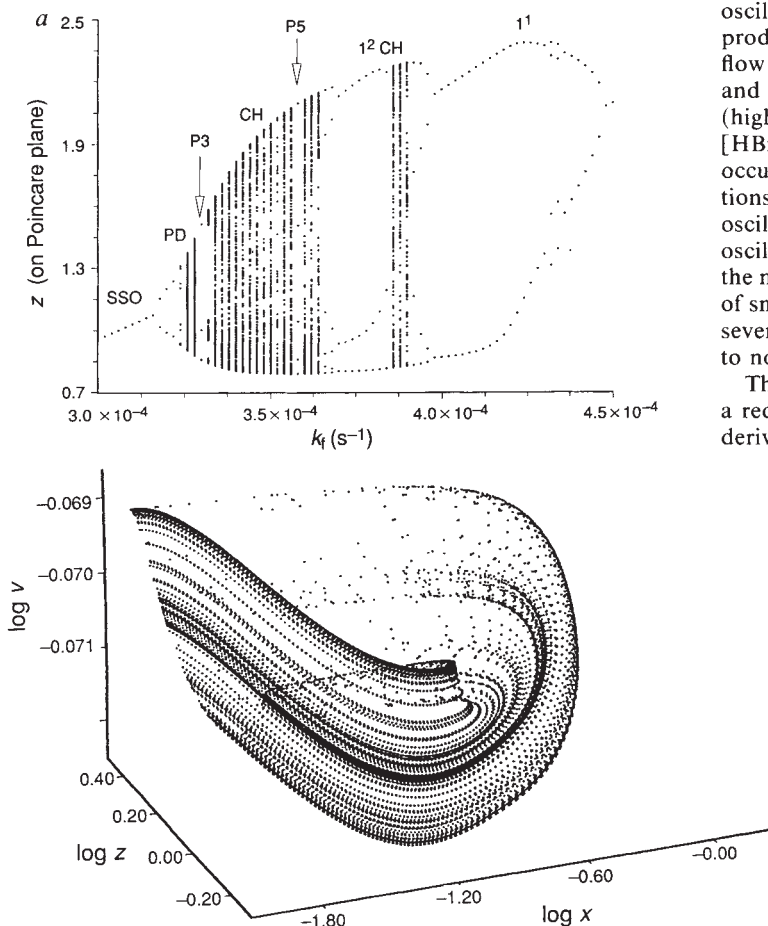


FIG. 1 *a*, Bifurcation diagram obtained from simulations based on the model in Table 2 at low CSTR flow rates with gradually increasing flow rate. The z ($[Ce(IV)]/[Ce(III)]_0$) values plotted on the ordinate are intersections of the trajectory with the Poincaré plane perpendicular to the x -axis and containing the point ($x=0.0468627$, $z=0.89870$, $v=0.846515$) when x is decreasing. Periodic behaviour is indicated when only a few intersections are visible at a particular k_t , whereas many intersections indicate chaos. Conditions for these simulations were: $A=0.1$ M, $M=0.25$ M, $H=0.26$ M, $C=0.000833$ M, $\alpha=666.7$ and $\beta=0.3478$. Notations used are: SSO, small-amplitude sinusoidal oscillations; PD, sequence of period-doubling bifurcations; P3, period-3 oscillations; P5, period-5 oscillations; CH, a chaotic regime; 1^0 and so on as described in the text. *b*, Three-dimensional view of the strange attractor obtained with the model in Table 2 at $k_t=3.9 \times 10^{-4} \text{ s}^{-1}$ with the conditions given in *a*. The symbols on the plot are separated by 2×10^{-4} dimensionless time units.

TABLE 1 Chemical scheme of a simple model of the BZ reaction

Reaction	Rates (r_i) and rate constants (k_i)
(1) $Y + X + H \rightarrow 2V$	$r_1 = k_1 H Y X$ $k_1 = 4.0 \times 10^6 \text{ M}^{-2} \text{ s}^{-1}$
(2) $Y + A + 2H \rightarrow V + X$	$r_2 = k_2 A H^2 Y$ $k_2 = 2.0 \text{ M}^{-3} \text{ s}^{-1}$
(3) $2X \rightarrow V$	$r_3 = k_3 X^2$ $k_3 = 3,000 \text{ M}^{-1} \text{ s}^{-1}$
(4) $0.5X + A + H \rightarrow X + Z$	$r_4 = k_4 A^{0.5} H^{1.5} (C - Z) X^{0.5}$ $k_4 = 55.2 \text{ M}^{-2.5} \text{ s}^{-1}$
(5) $X + Z \rightarrow 0.5X$	$r_5 = k_5 X Z$ $k_5 = 7,000 \text{ M}^{-1} \text{ s}^{-1}$
(6) $V + Z \rightarrow Y$	$r_6 = \alpha k_6 Z V$ $k_6 = 0.09 \text{ M}^{-1} \text{ s}^{-1}$
(7) $Z + M \rightarrow$	$r_7 = \beta k_7 M Z$ $k_7 = 0.23 \text{ M}^{-1} \text{ s}^{-1}$

Both the name and the concentration of chemical species are denoted by an uppercase letter. The concentrations of the main reactants (A, H and M) are fixed, as is the total concentration of cerium ions (C). The differential equations in Table 2 were obtained from this mechanism as described in the text. The chemical identities of the components are: $Y \equiv \text{Br}^-$, $X \equiv \text{HBrO}_2$, $Z \equiv \text{Ce(IV)}$, $V \equiv \text{BrCH(COOH)}_2$, $A \equiv \text{BrO}_3^-$, $H \equiv \text{H}^+$, $M \equiv \text{CH}_2(\text{COOH})_2$.

tion pathway via radical-transfer reactions, $\text{BrMA} \rightarrow \text{Br}^-$, for simplicity. This forces us to insert the adjustable parameters, α and β , multiplying the values⁹ of k_6 and k_7 respectively, to replace lost Br^- production and to account for the fact that products of the initial interaction of Ce(IV) with MA and BrMA go on to react with more Ce(IV) . The other rate constants reflect current knowledge of the BZ reaction^{9,13}. The model quite accurately reproduces the behaviour of the $\text{Ce(III)}-\text{BrO}_3^-$ reaction¹³

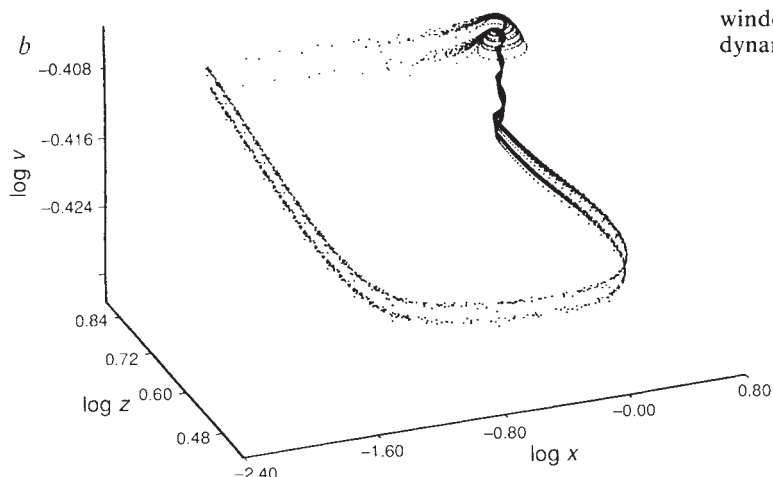
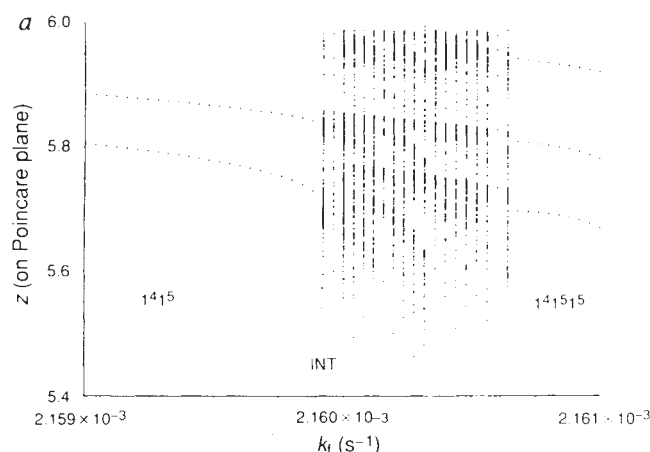


TABLE 2 Scaled differential equations

$$\begin{aligned} dx/d\tau &= T_0 \{-k_1 H Y_0 x \tilde{y} + k_2 A H^2 Y_0 / X_0 \tilde{y} - 2k_3 X_0 x^2 \\ &\quad + 0.5k_4 A^{0.5} H^{1.5} X_0^{0.5} (C - Z_0 z) x^{0.5} - 0.5k_5 Z_0 x z - k_7 x\} \\ dz/d\tau &= T_0 \{k_4 A^{0.5} H^{1.5} X_0^{0.5} (C/Z_0 - z) x^{0.5} - k_5 X_0 x z - \alpha k_6 V_0 z v \\ &\quad - \beta k_7 M z - k_7 z\} \\ dv/d\tau &= T_0 \{2k_1 H X_0 Y_0 / V_0 x \tilde{y} + k_2 A H^2 Y_0 / V_0 \tilde{y} + k_3 X_0^2 / V_0 x^2 \\ &\quad - \alpha k_6 Z_0 z v - k_7 v\} \end{aligned}$$

where
 $\tau \equiv t/T_0$, $x \equiv X/X_0$, $z \equiv Z/Z_0$, $v \equiv V/V_0$
 $\tilde{y} \equiv \{ \alpha k_6 Z_0 V_0 z v / (k_1 H X_0 x + k_2 A H^2 + k_7) \} / Y_0$

The equations describe deterministic chaos in the BZ reaction and result from the chemical mechanism in Table 1. Lower-case letters denote the dimensionless concentration of the chemical species. The scaling used in the above definitions is as follows: $T_0 = (10 k_2 A H C)^{-1}$, $X_0 = k_2 A H^2 / k_5$, $Y_0 = 4 k_2 A H^2 / k_5$, $Z_0 = C A / (40 M)$, $V_0 = 4 A H C / M^2$. The inverse residence time of the reactor (termed flow rate in the text) is k_f .

with $k_6 = k_7 = 0$. The mechanism also implements the observation¹³ that BrO_3^- and HBrO_2 usually are in equilibrium with BrO_2^+ during the autocatalytic $\text{BrO}_3^- + \text{Ce(III)}$ reaction. This results in a square-root dependence of the rate of reaction 4 on $[\text{HBrO}_2]$.

The initial set of differential equations (not shown) contains the four dimensionless variables $\tau = \text{time}/T_0$, $x = X/X_0$ ($X \equiv [\text{HBrO}_2]$), $y = Y/Y_0$ ($Y \equiv [\text{Br}^-]$), $z = Z/Z_0$ ($Z \equiv [\text{Ce(IV)}]$) and $v = V/V_0$ ($V \equiv [\text{BrMA}]$) resulting from the model in Table 1. Its dimension can be reduced by identifying variables that change on a faster timescale than the others^{14,15}. Either x or y may be eliminated by this criterion, but the resulting three-variable model is simpler and works better if $[\text{Br}^-]$ is eliminated by calculating $\tilde{y}$ as the root of the $dy/d\tau = 0$ equation. The remaining chemical species are HBrO_2 , Ce(IV) and BrMA . The resulting scaled differential equations are shown in Table 2.

We work with the original physical quantities rather than their scaled equivalents for clarity. The flow rate is used as bifurcation parameter with reactant concentrations identical to some used experimentally.

The feedstream concentrations at low CSTR flow rates, $A = [\text{BrO}_3^-] = 0.1 \text{ M}$, $M = [\text{MA}] = 0.25 \text{ M}$, $H = [\text{H}^+] = 0.26 \text{ M}$ and $C = [\text{Ce}]_{\text{tot}} = 0.000833 \text{ M}$, are the same as in the Texas experiments^{16,17}. We used the values $\alpha = 666.7$ and $\beta = 0.3478$. The precision to which α and β are quoted is a historical artefact; the results do not depend critically on these precise values (see refs 9–12 for justification of this procedure). Figure 1a shows the complete range of complexity at low flow rates with this parameterization. A reduced, low- $[\text{HBrO}_2]$ and low- $[\text{Ce(IV)}]$ steady state appears at $k_f \leq 2.0 \times 10^{-4} \text{ s}^{-1}$. Small-amplitude, sinusoidal oscillations develop as the flow rate is increased. A sequence of period-doubling bifurcations follows and results in chaos at $k_f \approx 3.2 \times 10^{-4} \text{ s}^{-1}$. This behaviour and the period-3 window at somewhat higher flow rates reproduce the universal dynamics¹⁸ of the logistic map. A period-5 window appears as

FIG. 2 a, Bifurcation diagram obtained from simulations based on the model in Table 2 at high CSTR flow rates with gradually increasing flow rate. The $z ([\text{Ce(IV)}]/[\text{Ce(IV)}]_0)$ values plotted on the ordinate are intersections of the trajectory with the Poincaré plane perpendicular to the x -axis and containing the point ($x = 0.446751$, $z = 5.75282$, $v = 0.393890$) when x is decreasing. Periodic behaviour is indicated when only few intersections are visible at a particular k_f , whereas many intersections indicate chaos. Conditions for these simulations were: $A = 0.14 \text{ M}$, $M = 0.3 \text{ M}$, $H = 0.26 \text{ M}$, $C = 0.001 \text{ M}$, $\alpha = 333.3$ and $\beta = 0.2609$. INT, intermittency route to chaos. b, Three-dimensional view of the strange attractor obtained with the model in Table 2 at $k_f = 2.16 \times 10^{-3} \text{ s}^{-1}$ with the conditions given in a. The symbols on the plot are separated by 2.5×10^{-3} dimensionless time units.

the flow rate is further increased, and after a narrow chaotic region, a cascade of reversed period-doubling bifurcations results in a 1^2 limit cycle. The small oscillations circle the low-[HBrO₂] and low-[Ce(IV)] steady state. Another chaotic window is found at even higher flow rates and is followed by a 1^1 periodicity. A chaotic attractor from this region is shown in Fig. 1b and is similar to experiment. Only isolated period-doubling bifurcations and periodic states were found at flow rates above the 1^1 periodicity. Large-amplitude, regular 1^0 oscillations appear above $k_f = 4.5 \times 10^{-4} \text{ s}^{-1}$ and persist until the bifurcation to the high-flow-rate steady state is approached, and the situation again becomes complex.

The high flow rate¹⁹ conditions are those of ref. 20: $A = [\text{BrO}_2] = 0.14 \text{ M}$, $M = [\text{MA}] = 0.3 \text{ M}$, $H = [\text{H}^+] = 0.26 \text{ M}$ and $C = [\text{Ce}]_{\text{tot}} = 0.001 \text{ M}$. We used $\alpha = 333.3$ and $\beta = 0.2609$. There are regular, 1^0 oscillations at $k_f = 2.02 \times 10^{-3} \text{ s}^{-1}$. But several complex periodic and chaotic windows appear as k_f is increased to $2.23 \times 10^{-3} \text{ s}^{-1}$ where the oxidized, high-[HBrO₂] and high-[Ce(IV)] steady state has been reached. The complex periodic oscillations consist of a mixture of small-amplitude sinusoidal cycles around the high-[HBrO₂] and high-[Ce(IV)] steady state and large-amplitude relaxation oscillations. The ratio of small to large oscillations increases as the flow rate increases, with the main periodic states being 1^1 , 1^2 , 1^3 and so on. Between any two of these periodic states there are more complex periodicities and chaos. The bifurcation diagram in Fig. 2a shows how the $1^4 1^5$ periodic window bifurcates through intermittency to chaos. The chaotic window contains several very complex periodic states and terminates in a $1^4 1^5$ limit cycle as the flow rate is increased. One of the strange attractors found in this region is shown in Fig. 2b. Small-amplitude, sinusoidal oscillations (0^1 periodicity) are found at $k_f = 2.22 \times 10^{-3} \text{ s}^{-1}$, just before reaching the high-[HBrO₂] and high-[Ce(IV)] steady state.

These simulations have many features in common with experiments. The complexity at both high and low flow rates is embedded between small-amplitude, sinusoidal oscillations and large-amplitude, 1^0 , relaxation oscillations. The presence and appearance of the low-flow-rate 1^1 and 1^2 periodicities and their period-doubling bifurcation are as found experimentally. The sequence and appearance at high flow rates of various complex limit cycles with chaotic windows between them is also common to experiment and simulations.

Two main differences between experiment and simulations exist at low flow rates. First, the experimental bifurcation leading from sinusoidal to mixed-mode oscillations is probably a secondary-Hopf bifurcation rather than the sequence of period-doubling bifurcations found in simulations. Second, there is no chaos between the 1^1 and the 1^0 periodicities in the simulations, even though most of the available chaotic experimental results were obtained in this region. We did not search for a set of parameters that might remove these discrepancies.

Although the main features of BZ-CSTR chaos can be rationalized using only homogeneous kinetics as is done here, it is clear that mixing effects are important¹¹. Chaos can be simulated by coupling mixing with a homogeneous model of the BZ kinetics that does not itself show chaos¹¹. But the detailed agreement between simulations based on the mixing model and experiment is not as good as that between experiment and simulations based on the 11-variable homogeneous model¹⁰. We are now investigating the effect of mixing on chaotic homogeneous models such as that presented here. □

Received 13 May; accepted 27 November 1991.

- Gleick, J. *Chaos. Making a New Science* (Viking, New York, 1987).
- Hao, B.-L. *Chaos II* (World Scientific, Singapore, 1990).
- Gray, P. & Scott, S. K. *Chemical Oscillations and Instabilities. Non-linear Chemical Kinetics* (Clarendon, Oxford, 1990).
- Field, R. J. & Burger, M. (eds) *Oscillations and Traveling Waves in Chemical Systems* (Wiley, New York, 1985).
- Belousov, B. P. in *Sbornik Referatov po Radiatsionnoi Meditsine* 145–147 (Medgiz, Moscow, 1958).
- Zhabotinsky, A. M. *Biophysika* 9, 306–311 (1964) (in Russian).

- Field, R. J. & Noyes, R. M. *J. chem. Phys.* 60, 1877–1884 (1974).
- Luo, Y. & Epstein, I. R. *Adv. chem. Phys.* 79, 269–299 (1990).
- Györgyi, L., Turányi, T. & Field, R. J. *J. phys. Chem.* 94, 7162–7170 (1990).
- Györgyi, L., Rempe, S. & Field, R. J. *J. phys. Chem.* 95, 3159–3165 (1991).
- Györgyi, L. & Field, R. J. *J. chem. Phys.* 91, 6131–6141 (1989).
- Györgyi, L. & Field, R. J. *J. phys. Chem.* 95, 6594–6602 (1991).
- Field, R. J. & Försterling, H.-D. *J. phys. Chem.* 90, 5400–5407.
- Thikhonov, A. N. *Math. Sb.* 31, 575–586 (1952) (in Russian).
- Tyson, J. J. *J. phys. Chem.* 86, 3006–3012 (1982).
- Roux, J.-C., Simoyi, R. H. & Swinney, H. L. *Physica D* 8, 257–266 (1983).
- Coffman, K. G., McCormick, W. D., Noszticzius, Z., Simoyi, R. H. & Swinney, H. L. *J. chem. Phys.* 86, 119–129 (1987).
- Feigenbaum, M. J. *J. stat. Phys.* 19, 25–52 (1978).
- Schmitz, R. A., Graziani, K. R. & Hudson, J. L. *J. chem. Phys.* 67, 3040–3044 (1977).
- Hudson, J. L. & Markin, J. C. *J. chem. Phys.* 74, 6171–6177 (1981).

Radiative forcing of climate from halocarbon-induced global stratospheric ozone loss

V. Ramaswamy*, M. D. Schwarzkopf† & K. P. Shine‡

* Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey 08542, USA

† NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey 08542, USA

‡ Department of Meteorology, University of Reading, Reading RG6 2AU, UK

OBSERVATIONS from satellite and ground-based instruments^{1–3} indicate that between 1979 and 1990 there have been statistically significant losses of ozone in the lower stratosphere of the middle to high latitudes in both hemispheres. Here we determine the radiative forcing of the surface-troposphere system^{4–6} due to the observed decadal ozone losses, and compare it with that due to the increased concentrations of the other main radiatively active gases (CO₂, CH₄, N₂O and chlorofluorocarbons) over the same time period. Our results indicate that a significant negative radiative forcing results from ozone losses in middle to high latitudes, in contrast to the positive forcing at all latitudes caused by the CFCs and other gases. As the anthropogenic emissions of CFCs and other halocarbons are thought to be largely responsible for the observed ozone depletions¹, our results suggest that the net decadal contribution of CFCs to the greenhouse climate forcing is substantially less than previously estimated.

The radiative forcing of the surface-troposphere system is central to concerns about climate change due to trace gases⁴. The direct radiative forcing is defined^{4,6} as the perturbation in the net radiative flux at the tropopause due to changes in concentrations of trace species, with water vapour changes and chemical interactions being excluded. This definition accounts for the rapid adjustment⁷ of stratospheric temperatures to a new equilibrium in response to radiative perturbations, while all tropospheric parameters are held fixed. The radiative forcings are determined by assuming that there is no change in the dynamical heating of the stratosphere, using the fixed dynamical heating (FDH) concept^{5,8–10}. New equilibrium temperatures are calculated for the stratosphere in response to the imposed radiative perturbations. The radiative fluxes calculated using the equilibrated stratospheric temperatures serve to determine the forcing⁴.

Two zonally averaged FDH models with fixed cloud amounts¹¹ are used here: the Geophysical Fluid Dynamics Laboratory (GFDL) 40-level model^{9,12} (latitude range 4.5 to 76.5° in 9° steps), and the 19-level University of Reading model^{13,14} (latitude range: 0 to 80° in 10° steps). Both models use a 10-cm⁻¹ narrow-band algorithm¹⁵ to calculate longwave radiation and assume cloud emissivities to be unity, but they have different solar radiation schemes, and different initial temperature, moisture and ozone climatologies. The GFDL results are for perturbations in January, April, July and October; the Reading results are for January and July in the Northern Hemisphere.

TABLE 1 Net radiative forcing of the surface-troposphere system

Latitude	Non-ozone gases		Ozone		
	CFCs (W m ⁻²)	All (W m ⁻²)	Solar radiative forcing (W m ⁻²)	Net radiative forcing without stratospheric temperature change (W m ⁻²)	Net radiative forcing with stratospheric temperature change (W m ⁻²)
90–60° N	0.05 (0.05)	0.30 (0.33)	0.12 (0.07)	0.04 (–0.06)	–0.15 (–0.18)
60–30° N	0.09 (0.08)	0.41 (0.43)	0.16 (0.11)	0.09 (0.01)	–0.13 (–0.15)
30–10° N	0.13 (0.14)	0.54 (0.53)	0.07 (0.08)	0.06 (0.04)	–0.04 (–0.04)
10° N–10° S	0.11 (0.12)	0.50 (0.51)	–0.01 (0.02)	–0.01 (0.01)	0.0 (–0.01)
10–30° S	0.12	0.53	0.0	0.0	0.0
30–60° S	0.10	0.43	0.19	0.11	–0.11
60–90° S	0.05	0.28	0.17	0.03	–0.30
Northern Hemisphere	0.10 (0.11)	0.45 (0.47)	0.10 (0.08)	0.06 (0.01)	–0.08 (–0.09)
Southern Hemisphere	0.10	0.46	0.10	0.05	–0.08
Global	0.10	0.45	0.10	0.06	–0.08

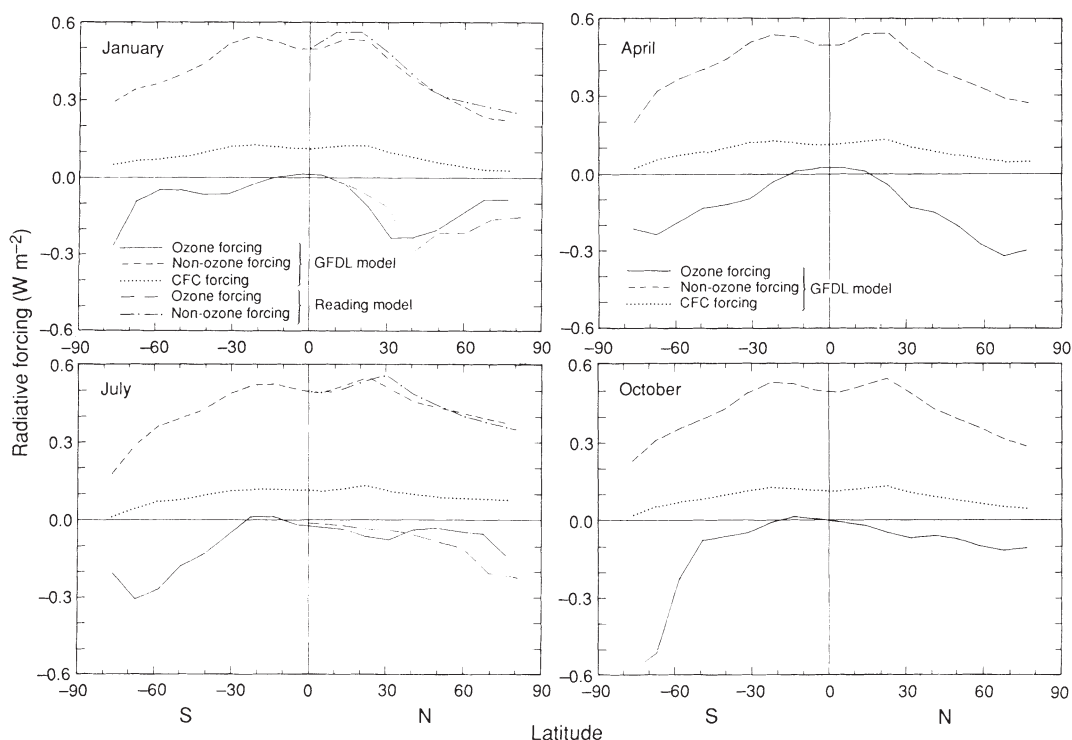
Net radiative (solar plus longwave) forcing due to the 1979–90 increases in CFCs and in all non-ozone gases, and that due to the observed stratospheric ozone losses. The net radiative forcing is obtained with the stratosphere in equilibrium in the presence of a fixed dynamical heating. The GFDL model results are averages over the four mid-season months, whereas the Reading model results (in parentheses) are averages over January and July only. Positive values denote gain of radiative energy.

The increases in the concentrations of the radiatively active, non-ozone gases between 1979 and 1990 are taken from ref. 1 (ch. 8). The changes in column ozone at each latitude over the same time period follow TOMS (total ozone mapping spectrometer) data². On the basis of the observed profiles^{1,3,16}, we assume that the depletions occur in the lower stratosphere, with an equal percentage change applied in each model layer between the tropopause and ~7 km above it (the vertical profile of the loss assumed here differs slightly from ref. 1). The location of the tropopause in the models is somewhat arbitrary and is used here to infer general implications for the surface-troposphere forcing. The tropopause in the GFDL model varies linearly with latitude from 100 mbar at the Equator to 270 mbar at 76.5°. In the Reading model, it is assumed to be at 96 mbar from 0 to 30° N (40° N in July), at 180 mbar from 30 to 50° N (60° N in July) and at 292 mbar thereafter.

The decadal increases in trace gases other than ozone induce a latitudinally dependent increase in the surface-troposphere forcing (Fig. 1 and Table 1). Loss of ozone in the lower stratos-

phere induces three distinct radiative effects^{5,6}: (1) more solar radiation reaches the surface-troposphere system (Table 1); (2) in the absence of stratospheric temperature changes, less longwave radiation is emitted into the surface-troposphere system and the ensuing forcing (Table 1) is lower than it would be for the solar effect alone; and (3) the *in situ* solar heating is reduced and there is a change in the convergence of longwave radiation in the lower stratosphere. Because the lower stratosphere is sensitive to radiative perturbations^{9,10,17,18}, this results, in the absence of any compensatory dynamical heating, in a decrease of temperatures at these altitudes. This, in turn, further reduces longwave emission from the stratosphere into the troposphere by an amount depending on the decrease in stratospheric temperature. The result in the middle to higher latitudes is an enhancement of the longwave effect over the solar, leading to a net negative radiative forcing (Table 1). Within the surface-troposphere system, the (positive) solar effect is 'felt' primarily at the surface, whereas the (negative) longwave effect is 'felt' primarily in the troposphere^{1,5}.

FIG. 1 Latitudinal dependence of the surface-troposphere radiative forcing due to the 1979–90 increases in (1) CFCs alone and (2) all the non-ozone gases (CO₂, CH₄, N₂O and CFCs), and that due to losses of ozone in the lower stratosphere. Results from the Reading model are for Northern Hemisphere January and July perturbations only, whereas the GFDL results are for perturbations corresponding to each of the four mid-season months (January, April, July and October). All the results were obtained by assuming stratospheric equilibrium in the presence of a fixed dynamical heating.



The resultant net ozone radiative forcing is compared with that due to the other radiatively active gases in Fig. 1 and Table 1. The ozone forcing depends on the amount of ozone loss² and insolation and thus on the season and latitude. Differences between the models arise because of different initial atmospheric profiles and different solar radiation algorithms. Poleward of 30°, the magnitude of the (negative) ozone forcings becomes comparable to and can even exceed the (positive) CFC forcing. At these latitudes, the magnitude of the ozone forcing can also be a considerable fraction of the (positive) non-ozone gas forcing. When globally and annually averaged (Table 1), the ozone forcing results in a substantial offset (~80%) of the decadal CFC greenhouse forcing, and a reduction of ~18% in the decadal greenhouse gas forcing.

There is strong evidence that the Antarctic ozone losses are primarily due to chlorine- and bromine-containing chemicals (halocarbons)¹. Evidence also suggests that a large part of the observed global loss in lower stratospheric ozone results from emissions of CFCs and other halocarbons¹. If this is so, then the net planetary greenhouse effect due to the halocarbons is the sum of their direct radiative effect and the indirect chemical effect by means of ozone depletion. Our calculations indicate that the chemical effects, in the global mean, substantially reduce the direct radiative effects of the CFCs. But the offset is not uniform with latitude (Fig. 1; Table 1); the total (CFCs plus ozone) forcing is positive at low latitudes owing to increases in CFCs, and negative at high latitudes owing to ozone losses. It would be difficult to decompose the observed ozone depletion into that due to individual halocarbons. Our calculations can only indicate the ozone-depleting effect of all halocarbons against the greenhouse effect of the CFCs.

Uncertainties in our calculations include cloud¹ and moisture distributions within the troposphere and the ozone trend³ below ~17 km. The radiative effects of stratospheric ozone are sensitive to the altitude profile of ozone loss¹⁹. Our calculations show that the magnitude of the ozone forcing would be reduced if the principal ozone losses were concentrated at altitudes higher than assumed here. For example, if the column ozone loss² were distributed on an equal percentage basis throughout the stratosphere, an assumption in which the principal losses occur at altitudes higher than observations indicate³, the GFDL and Reading models indicate a global mean forcing of -0.01 and -0.04 W m⁻² respectively. The ozone forcing could even become positive if the losses were confined to the upper stratosphere²⁰. In contrast to the stratospheric losses, a different set of factors¹ can cause an increase of ozone in the troposphere, yielding a positive forcing⁵. Because data are sparse the global trends due to this effect cannot be quantified at present¹. The need for accurate observations of changes in the total column and the vertical profile of ozone cannot be overemphasized.

We cannot yet ascertain whether the atmospheric dynamical response²¹ would enhance or reduce the ozone-induced radiative cooling tendency in the lower stratosphere. But general circulation model (GCM) studies of uniform perturbations of column ozone indicate that, unless the depletions are large (>50%), there is no significant change in the stratospheric dynamical activity²¹. The lower-stratospheric temperature response in these GCM studies is one of cooling and, in the middle and high latitudes, it is at least 70% of the response in the absence of dynamical changes^{9,10}. For an assumed scenario of global ozone changes and for the Antarctic springtime depletions, the GCM and FDH temperature responses are similar^{10,22,23}.

Observations for the periods 1973-87 (ref. 24) and 1958-89 (ref. 25) indicate a mean cooling trend for the global lower stratosphere (50-100 mbar). The temperature changes per decade estimated by two different linear-trend analyses in refs 24 and 25 are -0.34±0.38 K and -0.39±0.09 K, respectively, for the Northern Hemisphere, and -0.90±0.67 K and -0.45±0.15 K, respectively, for the Southern Hemisphere. The models give coolings per decade in the 50-100-mbar layer of -0.47 K

(GFDL; global) and -0.64 K (Reading; Northern Hemisphere), and occur mainly because of the ozone losses. Although the model results are within or at the higher end of the fairly large uncertainty limits in the observed trends, physical factors other than ozone losses^{21,26} (changes in tropospheric state, volcanic aerosols and so on) may also be contributing to the trends. Although we have investigated the radiative forcing of the surface-troposphere system, models of latitudinal changes in surface temperature and in the atmospheric general circulation require additional considerations (for example, the three-dimensional distribution of radiative perturbations^{6,27}, and dynamical feedbacks) and can be resolved only through GCM studies. □

Received 18 November 1991; accepted 7 January 1992.

1. WMO/UNEP *Scientific Assessment of Ozone Depletion: 1991* (WMO, Geneva, in the press).
2. Stolarski, R. S., Bloomfield, P., McPeters, R. D. & Herman, J. R. *Geophys. Res. Lett.* **18**, 1015-1018 (1991).
3. McCormick, M. P., Veiga, R. E. & Chu, W. P. *Geophys. Res. Lett.* (in the press).
4. Shine, K. P., Derwent, R. G., Wuebbles, D. J. & Morcrette, J.-J. in *Climate Change: The IPCC Scientific Assessment* (eds Houghton, J. T. et al.) 41-68 (Cambridge University Press, 1990).
5. Ramanathan, V. & Dickinson, R. E. *J. Atmos. Sci.* **36**, 1084-1104 (1979).
6. Ramanathan, V. et al. *Rev. Geophys.* **25**, 1441-1482 (1987).
7. Hansen, J. et al. *Science* **213**, 957-966 (1981).
8. Fels, S. B. & Kaplan, L. D. *J. Atmos. Sci.* **33**, 779-789 (1975).
9. Fels, S. B., Mahlman, J. D., Schwarzkopf, M. D. & Sinclair, R. W. *J. Atmos. Sci.* **37**, 2265-2297 (1980).
10. Kiehl, J. T. & Boville, B. A. *J. Atmos. Sci.* **45**, 1798-1817 (1988).
11. London, J. College of Engineering, New York Univ. *Rep. contract AF19(122)-165* (NTIS No. 117227) (1957).
12. Mahlman, J. D. & Umscheid, L. in *Dynamics of the Middle Atmosphere* (eds Holton, J. R. & Matsuno, T.) 501-525 (Terrapub, Tokyo, 1984).
13. Shine, K. P. *J. Atmos. Sci.* **48**, 1513-1518 (1991).
14. Slingo, A. & Schrecker, H. M. Q. *J. R. met. Soc.* **108**, 407-426 (1982).
15. Malkmus, W. *J. opt. Soc. Am.* **57**, 323-329 (1967).
16. Solomon, S. et al. *Nature* **321**, 755-758 (1986).
17. Kiehl, J. T. & Solomon, S. *J. Atmos. Sci.* **43**, 1525-1534 (1986).
18. Shine, K. P. Q. *J. R. met. Soc.* **113**, 603-633 (1987).
19. Laci, A. A., Wuebbles, D. J. & Logan, J. A. *J. geophys. Res.* **95**, 9971-9981 (1990).
20. Ramanathan, V., Cicerone, R. J., Singh, H. B. & Kiehl, J. T. *J. geophys. Res.* **90**, 5547-5566 (1985).
21. WMO *Report of the International Ozone Trends Panel 1988* ch. 6, *Global Ozone Research and Monitoring Project Rep. No. 18* (1989).
22. Kiehl, J. T., Boville, B. A. & Briegleb, B. P. *Nature* **332**, 501-504 (1988).
23. Shine, K. P. *Geophys. Res. Lett.* **13**, 1331-1334 (1986).
24. Angell, J. K. *J. Clim.* **1**, 1296-1313 (1988).
25. Oort, A. & Liu, H. *J. Clim.* (in the press).
26. WMO *Scientific Assessment of Stratospheric Ozone: 1989* ch. 1, *Global Ozone Research and Monitoring Project Rep. No. 20*, (1990).
27. Wang, W.-C., Dudek, M. P., Liang, X.-Z. & Kiehl, J. T. *Nature* **350**, 573-577 (1991).

ACKNOWLEDGEMENTS. K.P.S. acknowledges partial support from the UK Joint Environmental Program of National Power and Power Generation. We thank J. T. Kiehl, J. D. Mahlman, S. Manabe and V. Ramanathan for helpful discussions.

Upwelling and productivity changes inferred from a temperature record in the central equatorial Pacific

M. W. Lyle*, F. G. Prahl† & M. A. Sparrow†

* Borehole Research Group, Lamont-Doherty Geological Observatory, Palisades, New York 10964, USA

† College of Oceanography, Oregon State University, Corvallis, Oregon 97331-5503, USA

A SERIES OF C₃₇₋₃₉ alkenones synthesized by prymnesiophyte algae is commonly preserved in marine sediments, and can be used for estimating past sea surface temperatures (SSTs)^{1,2}. Here we present an alkenone analysis of sediments taken from the central equatorial Pacific Ocean (core W8402A-14GC; 0° 57' N, 138° 57' W) which shows that SST varied slightly (<2 °C) but coherently with Milankovitch insolation cycles over the past 250 kyr. The in-phase response of the SST to the precessional component of insolation indicates that this part of the SST time series may be driven by changes in local trade-wind strength, and the longer 100-kyr eccentricity component indicates a response to basin-wide South Pacific winds. Using a simple heat-balance

model, we use our palaeotemperature record to constrain the upwelling rates that have occurred in the past. Similar SST records from the eastern Pacific could constrain the horizontal advection component of the SST record and allow for better estimates of upwelling. Upwelling and palaeoproductivity then could be compared to determine whether changes in equatorial nutrient levels actually cause equivalent changes in productivity.

Sediment samples from the central equatorial Pacific core were extracted and analysed for alkenones as previously described³. Ages were assigned on the basis of the plankton $\delta^{18}\text{O}$ record of Murray⁴, shown in Fig. 1a. We used the index U_{37}^K , defined to quantify the degree of unsaturation in the alkenone biomarkers^{1,2}, to estimate SST for this time series. Laboratory and field experiments have shown that the index provides an accurate measure of absolute SST when the algae were growing. The alkenone SST estimate is not biased by diagenesis in deep sea sediments; comparison of the pristine centre of an organic-rich turbidite to its oxidized edge showed that 90% destruction of sedimentary organic carbon and

alkenones had no effect on the U_{37}^K SST estimate⁵. Analytical errors in measuring peak area limit the SST resolution to $\pm 0.4^\circ\text{C}$, on the basis of replicate analyses.

Here we aim to understand whether changes in equatorial Pacific SST result from changes in upwelling and, conversely, whether changes in upwelling respond directly to changes in tropical solar insolation or to changes in wind patterns driven by the appearance and disappearance of large Northern Hemisphere ice sheets. We also want to use the changes in SST to constrain the maximum changes in equatorial upwelling. In Fig. 1a, the SST time series from the central equatorial Pacific is compared to ETP (eccentricity, tilt, precession), a normalized and summed stack of the orbital forcing functions combined so that highest Northern Hemisphere insolation is represented by a positive displacement⁶. The equatorial SST record and orbital forcing are highly coherent (Fig. 1b). SST is coherent with the 100-kyr eccentricity cycle at $>95\%$ confidence interval and with the 23-kyr precession cycle at $>80\%$ confidence intervals. The coherence between the equatorial Pacific SST and the 41-kyr tilt cycle is slightly less than the 80% confidence interval. This coherence pattern is similar to orbital effects on tropical insolation. Tropical insolation responds much more strongly to eccentricity and precession than to tilt⁷.

High equatorial SST leads the minimum ice volume, based on the oxygen isotope record, by 5–10 kyr in all Milankovitch bands (compare, for example, the oxygen isotope signal and SST in Fig. 1a). In contrast, equatorial SST is in phase with Northern Hemisphere insolation (ETP signal) in the precessional frequency band and leads Northern Hemisphere insolation by ~ 13 kyr for eccentricity. The close match between equatorial Pacific SST and precessional Northern Hemisphere heating is probably associated with the expansion and contraction of the trade-wind belts. High equatorial SST probably occurs at times of weakened southeast tradewinds because higher austral winter insolation (June, July and August) and lower austral summer insolation weakens the temperature gradient that drives the winds⁷. Because the trade-wind belts are asymmetrical about the Equator, the average wind speed across the Equator weakens and the average position of the intertropical convergence zone moves southwards.

Cool SST in the eastern and central equatorial Pacific Ocean is primarily due to the dynamics of the Southern Hemisphere atmosphere and oceans. The trade winds are not symmetrical about the Equator but converge north of it, primarily because the 10–12°C average temperature difference between the Arctic and Antarctic results in an expanded high-pressure zone about the South Pole which displaces the westerlies and tradewinds northwards⁸. The southeast trade winds cross the Equator and cause upwelling. The large tongue of advected cold water in the eastern equatorial Pacific is also almost exclusively the result of Southern Hemisphere wind patterns. For this reason, the 10–15-kyr lead over eccentricity (and global ice cover) of the 100-kyr equatorial SST signal is probably due to a Southern Hemisphere influence, reflecting an asymmetry in the timing of climate change between the two hemispheres.

The amplitude of the SST time series can constrain the maximum level of equatorial upwelling associated with climate change. We constructed a simple box model for heat balance based on one for the modern central equatorial Pacific Ocean⁹. In the model, SST could only change by changes in upwelling. Solar heating was assumed constant throughout glacial and interglacial periods, an assumption which will hold true if changes in cloudiness are not a chief factor in tropical climate change. The temperature of upwelled waters and the horizontal heat advection into the equatorial box were also held constant. As SST reconstructions for the last glacial maximum¹⁰ have made it clear that the temperature of horizontally advected water was considerably lower in glacial periods, this model will always overestimate upwelling and thus must provide a maximum upwelling rate. With these constraints, the heat balance reduces

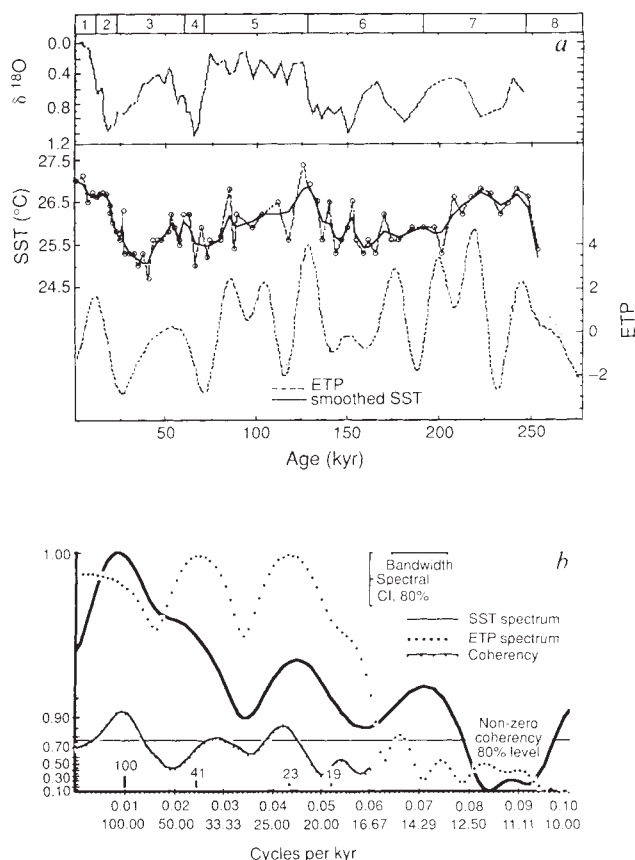


FIG. 1 a, The alkenone-estimated sea surface temperature (SST) time series for the past 250 kyr the central equatorial Pacific Ocean (core W8402A-14GC; $0^\circ 57' \text{N}$, $138^\circ 57' \text{W}$) compared with a standardized curve of Northern Hemisphere Milankovitch forcing (ETP)⁶. At the top of the figure is the $\delta^{18}\text{O}$ record from the planktonic foraminiferan *Globorotalia tumida* used to construct the timescale and the oxygen isotope stages. ($\delta^{18}\text{O} = [^{18}\text{O}/^{16}\text{O}]_{\text{sample}} / [^{18}\text{O}/^{16}\text{O}]_{\text{PDB}} - 1$). The alkenone SST, smoothed by a two-point moving average, is shown by a solid line, and raw data are shown connected by a dotted line. Highest SSTs are associated with highest Northern Hemisphere insolation. b, Cross-spectral analysis of ETP and the SST time series shows that the two series are highly coherent at the eccentricity (100-kyr) and precessional (23-kyr) Milankovitch bands. Amplitudes of the power spectra in each case have been scaled to the total range of the variance explained by each frequency band. Bandwidth denotes the frequency resolution of the spectra, and CI (confidence interval) denotes the size of peaks in the power spectrum with $>80\%$ probability of being significant.

to the following equation for the change in upwelling associated with a known change in temperature:

$$\Delta J_{up} = (T_0 - T_t) J_{out,0} / (T_t - T_{up})$$

where ΔJ_{up} is the change in upwelling flux, $J_{out,0}$ is the original horizontal advective flux out of the equatorial box, and T_0 , T_t , and T_{up} are the average temperatures for horizontally advected water at time 0 and at time t , and for the upwelled water.

In the modern central equatorial Pacific, ~15.3 Sv of water at 19.865 °C is upwelled and 41 Sv of water at 25.326 °C flows out of the region⁹. At 39 kyr we observe the minimum temperature in the SST time series, ~2° colder than the present; if this is due only to upwelling, it implies a rate ~2.5 times that of today (Fig. 2). For most of the past 250 kyr, upwelling rates must have less than 1.5 times the present rate. Thus, large changes in equatorial Pacific upwelling may have occurred in the past 250 kyr, but radical changes did not.

The important question, in terms of palaeoproductivity, is whether upwelling influences productivity change on glacial and interglacial timescales. Although upwelling provides nutrients needed for phytoplankton growth to the euphotic zone, there need not be a strong link between nutrients and productivity. For example, in the eastern equatorial Pacific there are regions with high surface concentrations of dissolved nitrate and silicate but relatively low primary productivity¹¹. Chavez *et al.*¹¹ argue that the high surface nutrient contents might be due to micro-nutrient limitation on phytoplankton growth, in particular, that of iron¹². The iron contained in upwelled water is sufficient for the observed regional primary productivity, but the nitrate should support roughly five times as much phytoplankton growth. If iron limits growth, then there should be an excess of other nutrients in equatorial surface waters. Addition of other sources of iron, such as new sources of aeolian dust¹², may cause productivity increases even in the absence of changes in upwelling.

Comparison of palaeoproductivity indices with palaeotemperature estimates will help determine whether upwelling or dust is the more important cause of changes in equatorial Pacific productivity. Additional SST time series in the Peru-Chile current region are needed, however, to monitor SST changes due to changes in horizontal advection. Even without this information, we can examine whether upwelling is sufficient to produce the palaeoproductivity record.

The modelled upwelling rates from the SST record are sufficient to have caused the change in productivity from the last

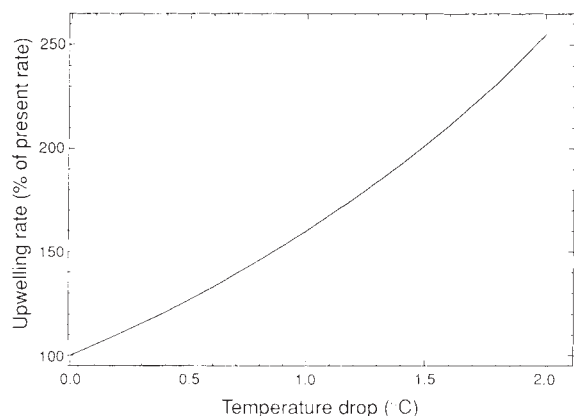


FIG. 2 Change in upwelling rate as a function of the temperature change from the modern central equatorial Pacific ocean⁹ (% of modern upwelling rate). Horizontal advection and solar insolation are assumed to remain constant through time, maximizing the changes in upwelling associated with a particular temperature drop. For the past 250 kyr, we have observed no more than a 2 °C temperature change in the equatorial Pacific, representing upwelling rates ≤ 2.5 times the present rates.

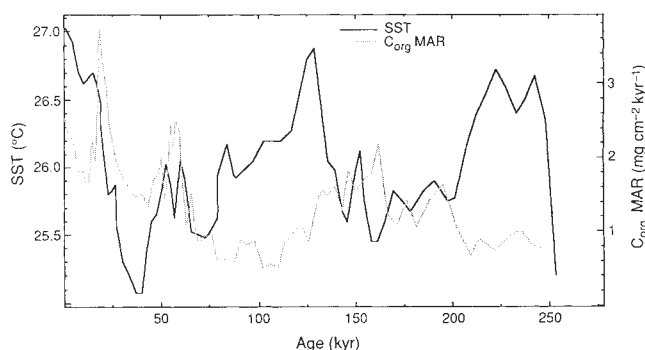


FIG. 3 Comparison between the organic carbon mass accumulation rate (MAR), an index of palaeoproductivity, and the alkenone-estimated SST record. Before 75 kyr, the two time series are negatively correlated, as would be expected if upwelling were a main factor controlling sea surface temperature. Between 70 and 20 kyr, horizontal advection, rather than upwelling, may be controlling the SST record.

glacial maximum (18 kyr) to the Holocene^{13,14}. New productivity in both the equatorial Atlantic and Pacific was ~50% higher at 18 kyr than today. The SST time series for W8402A-14GC indicates that central equatorial Pacific upwelling was no more than 20–50% higher at 18 kyr than at present, in agreement with other available data. But as the highest organic carbon burial in the sediments in the past 25 kyr occurs at a time of highest terrestrial (and presumably aeolian) input³, there is still a possibility that iron fertilization could have caused the change in productivity.

Comparing organic carbon burial rates^{15,16} and the SST record (Fig. 3) may help us to unravel whether the temperature changes were due to advection or upwelling. A horizontal advection event should transport relatively cold, nutrient-free water the central equatorial Pacific, as the water should have travelled at the surface in the euphotic zone for thousands of kilometres. An upwelling event should inject both nutrients and cold water into the region. Organic carbon burial is correlated with productivity, and may be an index of freshly upwelled waters^{13,15}. A period of enhanced upwelling should therefore produce a sedimentary interval with both cold SSTs and high organic carbon burial, whereas a period of enhanced advection of cold subpolar waters to the Equator should appear in the sedimentary record as cold SST and lower or unchanged organic carbon burial. Between 250 and 75 kyr, the SST and organic carbon mass accumulation rate are inversely correlated, as would be expected if the SST were controlled by upwelling events. The two records are nonetheless positively correlated for the interval between 70 and 20 kyr. Thus the cold interval between 40 and 20 kyr may represent the horizontal advection of cold water into the equatorial region.

Our detailed palaeotemperature record for the central equatorial Pacific Ocean, combined with other data from this core¹⁵, strongly constrains the regional upwelling rate and the maximum productivity changes in the region. Cross-spectral analysis with Milankovitch solar insolation cycles has indicated to us that the SST record responds rapidly to wind patterns in the Southern Hemisphere. It also provides working hypotheses for further testing. In particular, the equatorial SST record should have components reflecting both horizontal and vertical advection. By producing similar SST records from cores in the Peru-Chile Current near South America, and by comparing those records to that of the central equatorial Pacific, it should be possible to derive both of the upwelling and horizontal advection components through time. The refined upwelling record can then be used to understand equatorial productivity changes. □

Received 12 September; accepted 9 December 1991.

- Brassell, S. C., Eglinton, G., Mariow, I. T., Pflaumann, U. & Sarnthein, M. *Nature* **320**, 129–133 (1986).
- Prahl, F. G. & Wakeman, S. G. *Nature* **330**, 367–369 (1987).
- Prahl, F. G., Muehlhausen, L. A. & Lyle, M. *Paleoceanography* **4**, 495–510 (1989).
- Murray, D. W. thesis, Oregon State Univ. (1987).
- Prahl, F. G., deLange, G. J., Lyle, M. & Sparrow, M. A. *Nature* **341**, 434–437 (1989).
- Imbrie, J. *Milankovitch and Climate, Part I* (eds Berger, A. L. et al.) 269–305 (Reidel, Dordrecht, 1984).
- Berger, A. L. *Quat. Res.* **9**, 139–167 (1978).
- Flöhn, H. *Climate and Weather*, 111 (McGraw-Hill, New York, 1978).
- Bryden, H. L. & Brady, E. C. *J. phys. Ocean.* **15**, 1255–1273 (1985).
- CLIMAP Project Members *Geol. Soc. Am. Map and Chart Ser. MC-36* (1981).
- Chavez, F. P., Buck, K. R. & Barber, R. T. *Deep-Sea Res.* **37**, 1733–1752 (1990).
- Martin, J. H. *Paleoceanography* **5**, 1–13 (1990).
- Sarnthein, M., Winn, K., Duplessy, J. C. & Fontugne, M. R. *Paleoceanography* **3**, 361–399 (1988).
- Mix, A., *Nature* **335**, 541–544 (1989).
- Lyle, M. et al. *Paleoceanography* **3**, 39–59 (1988).
- Lyle, M. *Nature* **335**, 529–532 (1988).

Relationship between spreading rate and the seismic structure of mid-ocean ridges

G. M. Purdy*, L. S. L. Kong†‡, G. L. Christeson† & S. C. Solomon§

* Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

† WHOI/MIT Joint Program in Oceanography, Woods Hole, Massachusetts 02543, USA

§ Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

DIFFERENT parts of the global mid-ocean ridge system create oceanic crust at rates that differ by more than a factor of ten. The rates of supply of heat to these different systems must also vary substantially, and it is this heat supply, in the form of hot mantle material rising beneath the ridge axis, that determines the nature of the magmatic, tectonic and hydrothermal processes that control the production of new oceanic crust. Here we present evidence suggesting that a systematic relationship exists between the depth to zones of elevated temperature and partial melt beneath mid-ocean ridges and their spreading rate. Recent seismic measurements provide robust determinations at six different locations of the depth to the top of the axial low-seismic-velocity zone which suggest that as full spreading rate decreases from 150 to 20 mm yr⁻¹, the depth below the sea floor to the top of the LVZ increases from ~1 to ~4 km. Although this relationship might arise in a number of ways, our preferred explanation is that the major faults that extend to greater depths on slower-spreading ridges¹ allow a vigorous circulation of cooling sea water to reach greater depths, blocking the upward migration of large melt bodies.

The existence of substantial zones of elevated temperature and/or partial melt beneath mid-ocean ridges is inferred when seismic velocities are observed to decrease with depth beneath the axes of active mid-ocean ridge segments. Until recently, almost all the robust determinations of the depths to these low-velocity zones (LVZs) were located on the fast-spreading East Pacific Rise^{2–7}. New data from the Mid-Atlantic Ridge^{8,9}, the Juan de Fuca Ridge^{10,11} and the Lau Basin^{12,13} now permit a study of how the depth to these LVZs varies with spreading rate. A summary of these results is presented in Fig. 1. The Lau Basin determination is based on the identification of a bright intracrustal reflector which is inferred to be caused by a thin sill-like layer of magma¹³. The East Pacific Rise determinations are exceptionally well constrained by many refraction measurements that document LVZs^{2–7} and by extensive multichannel

seismic coverage of the axial magma chamber reflector^{14–17}. The southern East Pacific Rise datapoint is based on continuous along-axis reflection coverage of the rise axis between latitudes 14 and 20° S which clearly identifies a large-amplitude, flat-lying axial magma chamber reflector (ref. 18 and R. S. Detrick, personal communication). The remaining data points depend on

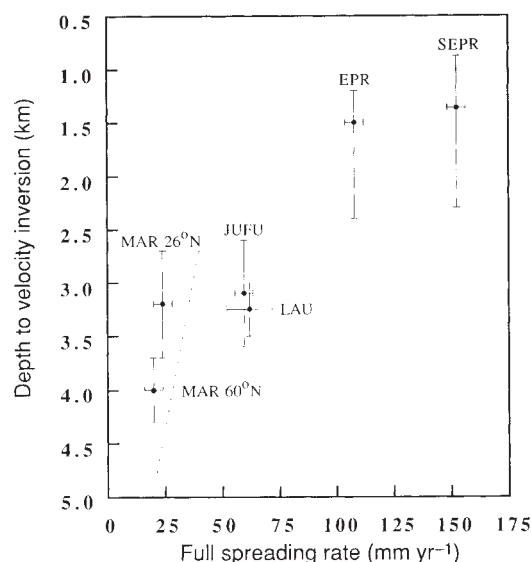


FIG. 1 Depth below sea floor to inversions in compressional wave velocity at six different spreading centres, plotted against full spreading rate. EPR: East Pacific Rise. Between 9° and 13° N, the roof of the magma chamber varies between 1.2 and 2.4 km below the sea floor, with an average depth of 1.5 km (ref. 33); this is the depth range shown. SEPR: East Pacific Rise south of the Garrett Fracture Zone between about 14 and 20° S where continuous along-axis CDP reflection profiles show that the depth to a prominent reflector associated with the axial magma chamber varies between 0.9 km and 2.3 km below the sea floor (ref. 18 and R. S. Detrick, personal communication). LAU: Valu Fa Ridge in the eastern Lau Basin^{12,13}. In ref. 13, amplitude analysis of wide-angle-reflection data shows a LVZ at a depth of 3 km, coincident with a bright reflector. Reference 12 has velocity control only from the limited aperture of a 2.4-km-long hydrophone array. The authors use depth migration to place the phase-shifted reflector at 3.5 km below the sea floor but suggest that their velocity estimates may be too high and thus their depth estimate too great. The depth range shown here is 3–3.5 km. JUFU: Northern Symmetric segment of the Juan de Fuca Ridge where new seismic refraction data suggest LVZ at an average depth of 3.2 km below the sea floor^{10,11}. Our best estimate of the uncertainty in this observation is ± 0.3 km. MAR 26°N: experiment near the TAG hydrothermal field on the Mid-Atlantic Ridge at latitude 26° N. Amplitude modelling and tomographic inversions of combined earthquake and explosive travel times constrain the depth of the LVZ to be 3 ± 0.5 km below the sea floor near an along-axis topographic high^{8,9}. MAR 60°N: careful analysis of explosive seismic refraction profiles³⁰ along the crest of the Reykjanes Ridge near latitude 60° N. Although the tectonic setting of these results is the most poorly known and the velocity inversion is small (0.2 km s^{-1}), the LVZ has considerable vertical extent (1.7 km) and is constrained by an exceptionally detailed analysis of the waveforms. The depth range of 3.7 to 4.3 km represents our subjective estimate of the uncertainty. The full spreading rates for SEPR¹⁸, EPR¹⁷, LAU³⁴, JUFU¹¹, MAR 26° N (ref. 9) and MAR 60° N (ref. 30) are respectively 152.5, 108, 60, 59.6, 24 and 20 mm yr⁻¹. Estimates of uncertainty in spreading rate are drawn as $\pm 4 \text{ mm yr}^{-1}$ except for LAU which is $\pm 10 \text{ mm yr}^{-1}$. The spreading rate of the Valu Fa Ridge is determined by modelling a small number of available magnetic anomaly profiles³⁴. The uncertainty is poorly constrained. Dotted line is an approximation to the maximum centroid depth of large ridge crest earthquakes plotted against spreading rate. This limiting relation, based on well constrained centroid depths of 35 mid-ocean ridge earthquakes (with focal mechanisms that are predominantly normal faulting on steeply dipping planes)¹, shows that the maximum centroid depth of ridge events decreases sharply with increasing spreading rate. The relationship is limited to spreading rates less than 40–45 mm yr⁻¹ because ridges spreading at higher rates rarely generate earthquakes of sufficient magnitude to be well recorded at teleseismic distances.

† Present address: Pacific Tsunami Warning Centre, 91-270 Fort Weaver Road, Ewa Beach, Hawaii 96706, USA

travel-time and amplitude interpretations of seismic refraction data that revealed the presence of velocity inversions.

The data set presented in Fig. 1 is small, but it suggests a relationship between spreading rate and the depth below the sea floor of substantial zones (on the seismic scale of ~ 1 km or more) of partial melt and/or elevated temperature¹⁹. If the LVZs observed beneath mid-ocean ridges are indeed related to high temperature and the presence of magma, as seems reasonable, then two plausible explanations exist for this relationship. It could be an artefact of temporal variability in the accretion process. Crustal emplacement, particularly at slow spreading rates, is not a steady-state process²⁰. It is reasonable to hypothesize that slowly spreading ridges pass through episodes of magmatic injection, separated by quiescent periods during which extensive cooling of the crustal column takes place. It is plausible that the time between these injection episodes increases with decreasing spreading rate. Seismic experiments on more slowly spreading ridges might then be more likely to be located over more completely cooled crust. The trend in Fig. 1 could then be dismissed as an interesting but unimportant artefact, even though our observations are already biased toward ridge segments that are in an active phase of their cycle, simply because we show results only from ridge segments that are underlain by a LVZ. No independent and reliable evidence exists to support this hypothesis, however, and it depends on chance location of the seismic measurements on more slowly spreading ridges.

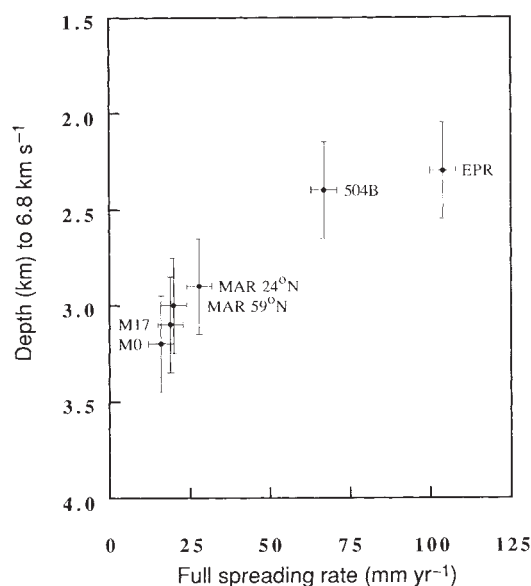


FIG. 2 Depth below the top of igneous basement to a velocity of 6.8 km s^{-1} (defined as the top of seismic layer 3 and nominally associated with the gabbroic layer) plotted against full spreading rate. Only high-quality data interpreted using rigorous travel-time inversion and amplitude modelling techniques are shown, and all experiments were done in regions with well-known tectonic frameworks, so interference from fracture zones, hotspots and other major structures is avoided. EPR is the East Pacific Rise at 13°N (ESP 12)⁵ where the depth to a velocity of 6.8 km s^{-1} is 2.3 km and the full spreading rate is 104 mm yr^{-1} (ref. 5), 504B is the region near drill site 504B on 6-Myr-old crust, 225 km south of the Costa Rica Rift (2.4 km and 67 mm yr^{-1})²⁸, MAR 24°N is located on 6-Myr-old crust south of the Kane Fracture Zone in the Central Atlantic Ocean (2.9 km and 28 mm yr^{-1})²⁷, MAR 59°N is on 9-Myr-old crust east of the Reykjanes Ridge near latitude 59°N (3.0 km and 20 mm yr^{-1})³⁰, M17 is in the Central Atlantic Ocean 300 km southwest of Bermuda on 140-Myr-old crust (3.1 km and 19 mm yr^{-1})³⁵, and M0 is in the same general area as M17 but over the 100-Myr-old crust of Mesozoic magnetic anomaly MO³¹ (3.2 km and 16 mm yr^{-1})³⁵. The discrepancies between the depths to the velocity inversions shown in Fig. 1 and the depths to the 6.8 km s^{-1} velocity contours (the presumed top of the frozen magma body) are explainable by regional or temporal variations in the accretion process or by the thickening of the upper crust by lava flows after it has moved away from the neovolcanic zone.

The second explanation is that Fig. 1 indeed demonstrates a systematic relationship between spreading rate and depth to substantial fraction of melt, as has been predicted by thermal models that include only conductive cooling^{21,22}. The thickness of the brittle crust and the maximum depth to which faulting extends beneath a mid-ocean ridge is controlled by poorly understood interactions between magmatic, tectonic and hydrothermal processes. We speculate that it depends on spreading rate because of the complex time-averaged balance that exists between rate of hydrothermal cooling (from the sea floor down) and the rate of supply of heat and magma (from the upper mantle into the crust). It is reasonable to assume that the rate at which hydrothermal circulation can remove heat increases as the depth of the heat source below the sea floor decreases. Therefore, the time-averaged balance between the rate of cooling and the heat supply is achieved at shallower depths beneath faster-spreading ridges where the heat supply, integrated over many thousands of years, is greater (simply because a larger volume of melt per unit time must be emplaced). Conclusions drawn from analyses of vent fluids from the Mid-Atlantic Ridge and the East Pacific Rise are not in agreement with this model²³, but these depend on the assumption that the silica equilibrium depth and the depth to the primary magma body are coincident.

Changes in structure with time can be expected, and changes in structure along axis are well documented^{7,9,11,24}. We consider these to be second-order effects compared with the primary variations related to spreading rate, but we expect that their magnitude will increase with decreasing spreading rate.

If we associate the top of the LVZ with the minimum depth reached by substantial volumes of material containing a significant fraction of melt, then we predict that the depth to the top of the gabbroic layer (commonly associated with seismic layer 3 and thus with velocities of 6.8 km s^{-1} and greater) increases with decreasing spreading rate. Previous models for the formation of oceanic crust have made this prediction²⁵, but it is difficult to test because the top of layer 3 can rarely be mapped in seismic experiments, and historical data on crustal velocity structures are dominated by variabilities related to factors other than spreading rate²⁶ (such as fracture zones, hotspots, or poor data and inadequate interpretation). Careful selection of high-quality determinations of seismic structure does show, however, that the preferred model is not inconsistent with observations: the depths to a velocity of 6.8 km s^{-1} show an increase of $\sim 1 \text{ km}$ as spreading rates decrease from 120 to 20 mm yr^{-1} (refs 5, 17, 27–31; Fig. 2). The number and distribution of data points in this figure are insufficient to establish that it is spreading rate alone that determines the depth to layer 3. We can conclude from Fig. 2 only that these selected data do not contradict the predictions that follow from the relationship depicted in Fig. 1.

Huang and Solomon¹ have shown that the maximum centroid depth of ridge-crest earthquakes (roughly half the maximum depth of brittle behaviour) increases from $2\text{--}3 \text{ km}$ at full spreading rates of $40\text{--}45 \text{ mm yr}^{-1}$ to $5\text{--}6 \text{ km}$ at full spreading rates of $5\text{--}10 \text{ mm yr}^{-1}$. This relationship (dotted line in Fig. 1) provides clear evidence that major faults extend deeper beneath the sea floor on more slowly spreading ridges. Both this relationship and that of the seismic structure data shown in Fig. 1 are too poorly constrained for significance to be attached to a comparison of their trends. The important coincidence is simply that both the depth to the velocity inversion and the maximum depth of faulting decrease with increasing spreading rate. These two observations could be related if faults are major conduits for the penetration of cooling water into the crust³², causing efficient cooling to greater depth beneath ridges with slower spreading rates. This greater tendency to cool the upper crust at slowly spreading ridges could block the upward migration of substantial volumes of melt and result in the relationship seen in Fig. 1.

We suggest that the minimum depth reached by significant volumes of melt along a ridge axis may be influenced by the maximum depth of penetration of major faults (and thus vigorous hydrothermal circulation) and that the consequent cooling constitutes an important mechanism for interaction between the magmatic and tectonic processes occurring on mid-ocean ridges. This relationship would not only lead to variations in crustal structure with spreading rate, but would also control second-order temporal and along-axis variations in the thickness of the volcanic layer. □

Received 30 September 1991; accepted 15 January 1992.

- Huang, P. Y. & Solomon, S. C. *J. geophys. Res.* **93**, 13445–13477 (1988).
- Orcutt, J. A., Kennett, B. L. N., Dorman, L. & Prothero, W. *Nature* **256**, 475–476 (1975).
- Rosendahl, B. R. et al. *J. geophys. Res.* **81**, 5244–5304 (1976).
- McClain, J. S., Orcutt, J. A. & Burnett, M. *J. geophys. Res.* **90**, 8627–8639 (1985).
- Harding, A. J. et al. *J. geophys. Res.* **94**, 12163–12196 (1989).
- Burnett, M. S., Caress, D. W. & Orcutt, J. A. *Nature* **339**, 206–208 (1989).
- Toomey, D. R., Purdy, G. M., Solomon, S. C. & Wilcock, W. S. D. *Nature* **347**, 639–645 (1990).
- Kong, L. S. L. thesis, Woods Hole Ocean. Inst., Massachusetts Inst. Technol. (1990).
- Kong, L. S. L., Solomon, S. C. & Purdy, G. M. *J. geophys. Res.* (in the press).
- Christeson, G. L., Purdy, G. M. & Rohr, K. M. *Eos* **70**, 1160 (1989).
- Christeson, G. L., Purdy, G. M. & Rohr, K. M. *Mar. geophys. Res.* (submitted).
- Morton, J. L. & Sleep, N. H. *Geology and Offshore Resources of Pacific Island Arcs—Tonga Region*, 441–453 (Circum-Pacific Council for Energy and Mineral Resources, Houston, 1985).

- Collier, J. & Sinha, M. *Nature* **346**, 646–648 (1990).
- Herron, T. J., Ludwig, T. J., Stoffa, P. L., Kan, T. K. & Buhl, P. *J. geophys. Res.* **83**, 798–804 (1978).
- Detrick, R. S. et al. *Nature* **326**, 35–41 (1987).
- Mutter, J. C. et al. *Nature* **336**, 156–158 (1988).
- Vera, E. E. et al. *J. geophys. Res.* **95**, 15529–15556 (1990).
- Detrick, R. S. et al. *Eos* **71**, 506 (1991).
- Purdy, G. M., Kong, L. S. L., Christeson, G. L. & Solomon, S. C. *Eos* **71**, 627 (1990).
- Ballard, R. D. & van Andel, T. H. *Geol. Soc. Am. Bull.* **88**, 507–530 (1977).
- Sleep, N. H. *J. geophys. Res.* **80**, 4037–4042 (1975).
- Royden, L. & Parsons, B. *Geophys. J. R. astr. Soc.* **70**, 741–753 (1982).
- Campbell, A. C. et al. *Nature* **335**, 514–519 (1988).
- Purdy, G. M. & Detrick, R. S. *J. geophys. Res.* **91**, 3739–3762 (1986).
- Cann, J. R. *Geophys. J. R. astr. Soc.* **39**, 169–187 (1974).
- Purdy, G. M. & Ewing, J. I. in *The Geology of North America*, Vol. M (eds Tucholke, B. E. & Vogt, P. R.) 313–330 (Geological Society of America, Boulder, 1986).
- Detrick, R. S. & Purdy, G. M. *J. geophys. Res.* **85**, 3759–3777 (1980).
- Collins, J. A., Purdy, G. M. & Brocher, T. M. *J. geophys. Res.* **94**, 9283–9302 (1989).
- Purdy, G. M. *Geophys. J. R. astr. Soc.* **72**, 115–137 (1983).
- Bunch, A. W. H. & Kennett, B. L. N. *Geophys. J. R. astr. Soc.* **61**, 141–166 (1980).
- Mithal, R. & Mutter, J. *Geophys. J.* **97**, 275–294 (1989).
- Cann, J. R., Strens, M. R. & Rice, A. *Earth planet. Sci. Lett.* **76**, 123–134 (1985).
- Detrick, R. S. in *Ophiolite Genesis and Evolution of the Oceanic Lithosphere* (eds Peters, T. et al.) 7–20 (Ministry of Petroleum and Minerals, Sultanate of Oman, 1991).
- Morton, J. L. & Pohl, W. *Geologisches Jahrbuch* (eds von Stackelberg, U. & von Rad, U.) 93–108 (1990).
- Klitgord, K. D. & Schouten, H. in *The Geology of North America*, Vol. M (eds Vogt, P. R. & Tucholke, B. E.) 351–378 (Geological Society of America, Boulder, 1986).

ACKNOWLEDGEMENTS. We thank J. Collins and D. DuBois for help in compiling these results and J. Lin for discussions. This research was supported by the Office of Naval Research and the NSF.

Evolution of ecological differences in the Old World leaf warblers

Adam D. Richman* & Trevor Price

Department of Biology O-116, University of California at San Diego, La Jolla, California 92093, USA

SYMPATRIC species that belong to the same ecological guild usually differ in their behaviour and morphology, and these differences are often interpreted as adaptations to having to make use of different resources. Evidence supporting this interpretation comes from association between ecology and morphology among species¹, in which an *a priori* functional relationship is reasonable. But one problem with such comparisons is that members of a guild may be closely related, so the more closely related species can share a greater similarity in their morphology and ecology simply as a result of the lingering legacy of a common ancestor^{2–5}. In principle, the importance of historical legacy can be evaluated from phylogenetic relationships and times since divergence for all species^{2,6}, but this is rarely possible because these data are not available. Here we use a phylogeny for eight sympatric species of warbler in the genus *Phylloscopus*, based on their mitochondrial DNA sequences, to remove the effects of historical legacy. Without these effects, we find strong support for adaptive interpretations of among-species variation in habitat selection, prey-size choice and feeding method. Ecological variation along any of these three niche axes is associated with predictable morphological variation. We also find evidence for historical legacy in that more closely related species are often more similar behaviourally and morphologically. This paradoxical result can be reconciled because the most closely related species tend to differ along only one niche axis, habitat choice. In contrast, the evolution of prey-size choice and feeding method occurred rapidly and early in the diversification of this group. Once a new ecological zone was occupied, subsequent morphological change along these niche axes was limited, accounting for the similarity of closely related species.

We have studied the ecology and breeding of eight species of warblers, which occur in sympatry during the breeding season in the Himalayas of Kashmir, India. The eight warbler species all belong to the same genus (*Phylloscopus*), and together with the goldcrest, *Regulus regulus*, they form the insectivorous leaf-gleaning guild⁷. The species have been shown to differ in habitat choice, size of prey, and feeding method⁸. There are strong correlates of morphology with the ecological differences⁸, as illustrated in Fig. 1a, c and e. Larger species feed on larger prey, species with relatively short tarsi breed in coniferous (versus deciduous) woodlands, and species with wide beaks tend to flycatch more.

We estimated the phylogeny for 19 specific and subspecific taxa within the *Phylloscopus* plus a single species from three other warbler genera, including *Regulus*. The data consist of a contiguous DNA sequence of 910 bases within the mitochondrial cytochrome *b* gene (Fig. 2a). Most observed variation among these closely related species is at third-position sites. The number of observed transitions among species of *Phylloscopus* at third-position sites is similar to that observed for more distant inter-generic comparisons, indicating rapid saturation for transitional changes within the genus, whereas the number of the more slowly accumulating (observed) transversion differences is not similarly affected (Fig. 2b). Similar results have been found in other mitochondrial DNA sequence studies of closely related species^{9–11}. Moreover, a study of cytochrome *b* evolution in mammals found that transversions continue to accumulate approximately linearly for tens of millions of years¹¹. Corrections for multiple substitutions at third-site transversions yield pairwise distances little different from those observed (data not shown). We therefore used the pairwise fraction of observed (untransformed) transversion differences over all sites as a linear measure of species divergence, and estimated the phylogeny using the neighbour-joining algorithm¹² implemented in the program NEIGHBOR in PHYLIP¹³.

The neighbour-joining tree for 22 taxa is shown in Fig. 3. Taxonomic groups identified by the molecular phylogeny are in good agreement with previous suggestions^{14,15} based on characters such as nasal bristles and plumage patterns which were used to distinguish major groups (see Fig. 3). We tested the sensitivity of the tree topology to sampling variation using the bootstrap¹⁶. The frequencies of occurrence of different groups within the eight Indian *Phylloscopus* at our study site in

* Present address: Department of Biology, University of Oregon, Eugene, Oregon 97403-1210, USA.

Kashmir, and the three outgroup species, were extracted from bootstrap replicates using the larger species set, and are shown in Table 1. The existence of three major clades within this group of *Phylloscopus* is reasonably well supported, with somewhat less support for the order of branching within and among these clades.

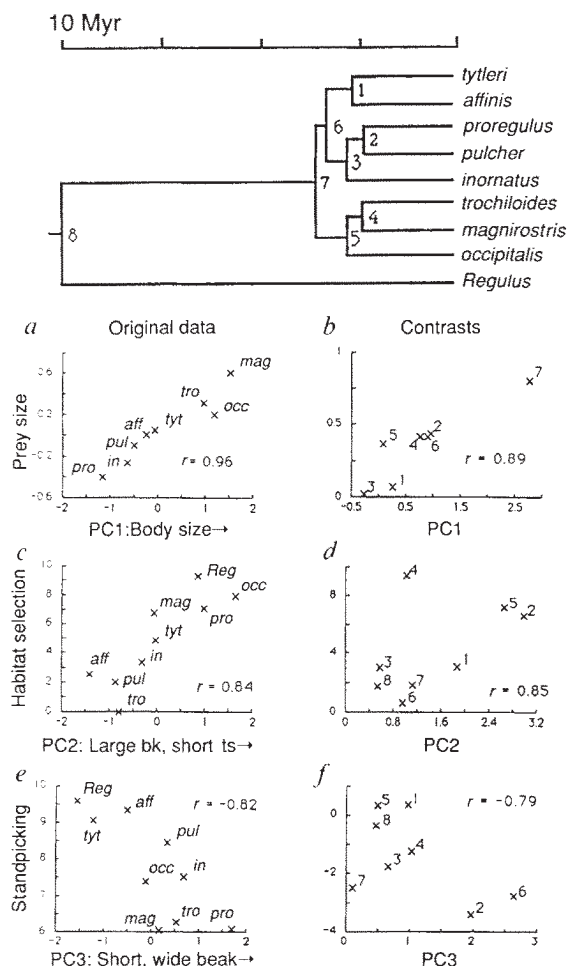


FIG. 1 Top. The phylogeny for eight species of *Phylloscopus*, and *Regulus*, where branch lengths are proportional to time. Branch lengths generated by the program NEIGHBOR do not assume rate constancy, and so are not necessarily proportional to time. Branch lengths were obtained for the topology in Fig. 3 using DNAMLK¹³, a maximum likelihood method which assumes a molecular clock. Only third position sites were used to estimate branch lengths: both transitions and transversions were used, and the (maximum likelihood) ratio of transitions to transversions was specified as 5:1. Numbered tree nodes refer to specific contrasts between taxa. For example, contrast 4 compares {*magnirostris*} to {*trochiloides*}, and contrast 6 compares {*pulcher*, *proregulus*}, *inornatus*} to {*affinis*, *tytleri*}. Upper scale represents a calibration of branch length to time, using an estimate of time since divergence between *Phylloscopus* and *Regulus* according to DNA-DNA hybridization data¹⁹. a-f, Original correlations and contrast correlations (using the tree and branch-lengths presented). Morphology is along the abscissa. Principal components were extracted from the correlation matrix of mean values of ln-transformed species⁸. PC1 is a summary measure of body size. PC2 contrasts beak (bk) size to tarsus length (ts) and PC3 is a measure of beak shape. Prey size refers to beetle head widths in faecal analysis: faeces for *Regulus* were not available. Habitat selection is the (transformed) percentage of feeding observations in coniferous vegetation, and is highly negatively correlated with altitude. Stand-picking measures the (transformed) percentage of feeding observations where capture of prey did not involve flight (see ref. 8 for details). Contrast correlations were calculated assuming the average of all contrasts for each character is zero^{4,20}, so correlations for the original data and the corresponding contrast data are not directly comparable. All correlations are significant ($P < 0.05$).

The importance of historical legacy was evaluated using Felsenstein's method of independent contrasts, which requires information on phylogenetic relationships and times since divergence for all species². The method assumes that resemblance between species due to historical legacy decreases with time as the result of the accumulation of many small changes, so that it is adequately described by a Brownian motion model. Large coordinate changes in morphology and ecology are evidence against a purely historical hypothesis. Only pairs of species whose evolutionary histories are independent of all other species in the tree are examined; for example, all sister pairs are considered. These species are then removed from the tree, and their common ancestor is treated as a new species with morphological and ecological character values that are a weighted average of the two descendants². The pruning procedure creates $(n-1)$ independent contrasts (from the n species) in terms of the observed amount of character change between two taxa, scaled by the length of time since their divergence. A correlation between the morphological and the ecological contrasts, which we term a 'contrast correlation', cannot be due to historical legacy and supports an adaptive explanation.

In Fig. 1b, d and f we show scatterplots of morphological and ecological contrasts as constructed using Felsenstein's method² for each of the corresponding morphology/ecology scatterplots (Fig. 1a, c, e, respectively). The contrast analyses confirm the associations between morphology and ecology, once history is removed. This result is surprising in view of the fact that historical legacy is often evident in these data. For example, the effect of phylogeny on body size in Fig. 1a is clear: the three smallest *Phylloscopus* species are most closely related, as are the three largest, and the two intermediate size species. Nevertheless, once this non-independence of morphology and ecology, once history is removed, body size remains correlated with prey size. How is this possible? One reason is that the same trend is repeated within each of the size clades (Fig. 1b, contrasts 1, 2, 4). But the result does not depend on the almost perfect rank correlation in the original data. The largest contrast is between small and large body size clades (Fig. 1b, contrast 7), and this contrast is robust to changes in rank order within clades. Contrast 7 (Fig. 1b) is large not only because it compares (ancestral) taxa of very different size, but because the size shift occurred over a relatively short time, as measured by the short interior branch lengths.

A second result is that different correlations are upheld for different reasons. In the first plot (Fig. 1b), the greatest contrast was between large- and small-bodied clades, but this contrast is fairly small in other correlations. In the second plot (Fig. 1d), internal contrasts of morphology (PC2) between major clades are relatively small (contrasts 6, 7, 8) and large differences in habitat and/or related morphology occur between pairs of species within each of the size clades (contrasts 1, 2, 4, 5). In the third plot (Fig. 1f), there are large contrasts in beak shape and related foraging behaviour both in a pair of sister species (contrast 2) and in an internal contrast (contrast 6).

We found little effect of uncertainty in the reconstruction of phylogeny on the contrast correlations. We repeated the analysis using 100 neighbour-joining trees found from bootstrapped data sets, assuming *Regulus* was the outgroup. Correlations were 0.89 ± 0.02 s.d., 0.86 ± 0.02 s.d. and -0.80 ± 0.03 s.d. for the three sets of contrasts in Fig. 1b, d and f.

Using the results of the contrast analyses in conjunction with the phylogeny, we can reconstruct the temporal sequence of evolution of ecological segregation within the *Phylloscopus*. The first split within the *Phylloscopus* is accompanied by a large change in body size, as indicated by the large value of contrast 7 (Fig. 1b). Subsequent changes in body size within each of these size clades are small compared with the inferred change from the common ancestor to both groups. A secondary splitting is accompanied by a large change in feeding morphology and related behaviour (Fig. 1f, contrast 6). *P. affinis* and *P. tytleri* are mainly standpickers, and have relatively narrow, long beaks

A

[illegible]

by P. Berthold, Max Planck Institute; *Regulus* material by C. Wood, University of Washington Museum. Amplifications were done in 50 μ l Tris (67 mM, pH 8.8) containing 1.5 mM MgCl₂, with 1 mM each dNTP, 1 μ M each primer, template DNA (0.1–1 μ g) and 1–2U *Taq* polymerase. Primers used for amplification and sequencing were L14841 and H15149 (ref. 10), L15752 (5'-TACTGTTGGTACCACCGATTCAAGT-3'), L15055 (5'-CACATCGGCCGAGGATTCTAC-3'), L15532 (5'-GGGTGGAATGGGATTTT-3'), L15557 (5'-AATAGGAAGTATCATTC-3'), and H15629 (5'-TGAGGAATAGGACTAGCA-3'). The temperature profile for 30 cycles was 1 min at 92 °C, 1 min at 50 °C, and 2 min at 72 °C. Purification in 2.5% agarose of the initial amplification product was followed either by single-strand amplification of each strand and sequencing using the limiting primer, or cloning into Bluescript vector and subsequent dideoxy sequencing. Three clones for each sequence were isolated for sequencing.

compared with all other species of *Phylloscopus*. Thus, an early ecological innovation seems to have been a change in feeding method. Finally, large differences between more closely related species are observed in habitat choice and/or associated morphology (Fig. 1d, contrasts 1, 2, 4, 5). Thus, early ecological innovations are changes in feeding method and prey size, with habitat selection more recently modified. The inferred sequence of the evolution of ecological differences is reasonably robust to possible errors in the phylogeny, because it depends largely on the monophyly of the three major clades, which occur with high frequency in bootstrap analyses (Table 1).

The approximate timescale in Fig. 1 indicates that all these *Phylloscopus* species may have originated before the beginning of the Pleistocene. Rapid speciation and diversification occurred very early in the history of this group and was not driven by repeated periods of climatic change and subsequent restriction of forest habitats, which are known to have occurred¹⁷. Closely related species mostly differ in habitat and related morphology, yet it appears that these species pairs have been in existence for millions of years. The pattern may be attributed to recent geographic contact¹⁸, due to spread of forests. Alternatively, the present ecological community may be a stable entity which has persisted largely unchanged over long periods of time.

Two results from our study of the evolution of ecological differences among *Phylloscopus* warblers are of general significance. First, the adaptive interpretation of morphological and behavioural variation within this group is upheld, despite

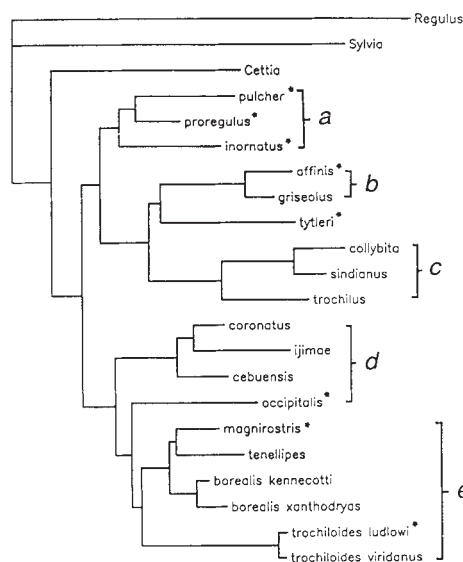


FIG. 3 The neighbour-joining tree for 22 taxa, obtained using the fraction of observed transversion differences over all sites. Asterisks denote the *Phylloscopus* species that occurred at our study site in Kashmir. Lettered brackets to the right indicate taxonomic groups previously recognized on the basis of morphological data^{14,15}. a, Dusky warblers; b, European warblers; c, Asiatic warblers; d, greenish warblers; e, crowned warblers (*cebuensis* was not considered by Williamson¹⁴); f, proposed subgenus *Acanthopneuste*¹⁵. There is a large degree of concordance between the molecular phylogeny and previous classifications (which did not consider relationships within or among groups of more closely related species). The two exceptions are: (1) the crowned warblers (group d) are not monophyletic according to the molecular phylogeny, with the western crowned warbler *P. occipitalis* placed as the outgroup to the greenish warblers (group e), and (2) the position of *tytleri* in the molecular phylogeny is consistent with the classification of Ticehurst¹⁵, but not with that of Williamson¹⁴. Williamson grouped *tytleri* with the greenish warblers (group e). Williamson's classification in this instance is not supported by the bootstrap data, where *tytleri* never occurs with the greenish warblers (Table 1).

TABLE 1 Unrooted groups occurring at least five times in 100 bootstrap replicates*

{ <i>magnirostris</i> , <i>trochiloides</i> , <i>occipitalis</i> }†	88
{ <i>affinis</i> , <i>tytleri</i> }†	86
{ <i>proregulus</i> , <i>pulcher</i> , <i>inornatus</i> }†	71
{ <i>Regulus</i> , <i>Sylvia</i> , <i>Cettia</i> }†	68
{ <i>Regulus</i> , <i>Sylvia</i> }†	63
{ <i>proregulus</i> , <i>pulcher</i> }†	63
{ <i>magnirostris</i> , <i>trochiloides</i> }†	53
{ <i>affinis</i> , <i>tytleri</i> , <i>proregulus</i> , <i>pulcher</i> , <i>inornatus</i> }†	48
{ <i>magnirostris</i> , <i>occipitalis</i> }	39
{ <i>proregulus</i> , <i>inornatus</i> }	34
{ <i>Regulus</i> , <i>Sylvia</i> , <i>Cettia</i> , <i>affinis</i> , <i>tytleri</i> }	30
{ <i>Regulus</i> , <i>Cettia</i> }	20
{ <i>Regulus</i> , <i>Sylvia</i> , <i>affinis</i> , <i>tytleri</i> }	16
{ <i>Cettia</i> , <i>magnirostris</i> , <i>trochiloides</i> , <i>occipitalis</i> }	13
{ <i>Regulus</i> , <i>affinis</i> , <i>tytleri</i> }	10
{ <i>Sylvia</i> , <i>Cettia</i> }	9
{ <i>inornatus</i> , <i>affinis</i> , <i>tytleri</i> }	6
{ <i>Cettia</i> , <i>Sylvia</i> , <i>magnirostris</i> , <i>trochiloides</i> , <i>occipitalis</i> }	6
{ <i>magnirostris</i> , <i>trochiloides</i> , <i>occipitalis</i> , <i>inornatus</i> }	6
{ <i>occipitalis</i> , <i>proregulus</i> , <i>pulcher</i> , <i>inornatus</i> , <i>affinis</i> , <i>tytleri</i> }	5
{ <i>proregulus</i> , <i>pulcher</i> , <i>affinis</i> , <i>tytleri</i> }	5

* Using the neighbour-joining program NEIGHBOR in PHYLIP¹³.

† Indicates that the group is also found on the neighbour-joining tree. In the bootstrap resampling procedure, first and second sites, and third sites alone, were resampled separately and then combined for phylogenetic analysis.

evidence for the importance of phylogeny in causing resemblance among *Phylloscopus* species. The explanation for these apparently contradictory findings is that closely related species tend to differ along a single morphological axis associated with habitat selection, with usually relatively little change along other niche axes. Rather than providing support for the common operating assumption of no phylogenetic constraint, these results underscore the argument² that phylogenies are fundamental to comparative biology. Second, the use of a well supported phylogeny in conjunction with the method of independent contrasts provides insight into the mode and tempo of character change. The sequence of evolution of ecological characters reported, namely that habitat differentiation has changed most recently, and is therefore probably the first step in species divergence, followed by changes in diet, has been shown for other groups using comparisons among genera¹⁸. Thus, it may be a normal mode of species divergence. The contrasts (Fig. 1b, d, f) suggest that there have been long periods of little ecological and morphological change and occasional periods of quite rapid change, usually in just one niche dimension. We believe the observed resemblance between ancestral and descendant taxa within the *Phylloscopus* says little about limitations in the ability of species to respond to selection due to genetic constraints. Instead, it reflects changes in the adaptive topography as the diversity of sympatric species increased over evolutionary time. □

Received 6 November; accepted 11 December 1991.

1. Miles, D. B. & Ricklefs, R. E. *Ecology* **65**, 1629–1640 (1984).
2. Felsenstein, J. *Am. Nat.* **125**, 1–15 (1985).
3. Harvey, P. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, UK, 1991).
4. Harvey, P. & Purvis, A. *Nature* **351**, 619–624 (1991).
5. Brooks, D. R. & McLennan, D. A. *Phylogeny, Ecology, and Behavior* (Univ. Chicago Press, 1991).
6. Lynch, M. *Evolution* **45**, 1065–1080 (1991).
7. Price, T. & Jamdar, N. *J. Bombay Nat. Hist. Soc.* **87**, 1–15 (1990).
8. Price, T. *J. Animal Ecol.* **60**, 643–664 (1991).
9. Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. *J. molec. Evol.* **18**, 225–239 (1982).
10. Kocher, T. et al. *Proc. natn. Acad. Sci. U.S.A.* **86**, 6196–6200 (1989).
11. Irwin, D. M., Kocher, T. D. & Wilson, A. C. *J. molec. Evol.* **32**, 128–144 (1991).
12. Saitou, N. & Nei, M. *Molec. biol. Evol.* **4**, 406–425 (1987).
13. Felsenstein, J. *PHYLIP Manual Version 3.4* (Herbarium, Univ. California, Berkeley, 1991).
14. Williamson, K. *The Genus Phylloscopus* (British Trust for Ornithology, Oxford, 1974).

15. Ticehurst, C. B. *A Systematic Review of the Genus Phylloscopus* 1–193 (Trustees of the British Museum, London, 1938).
16. Felsenstein, J. *J. molec. Evol.* **17**, 368–376 (1981).
17. Frenzel, B. *Science* **161**, 637–649 (1968).
18. Diamond, J. in *Community Ecology* (eds Diamond, J. & Case, T. J.) 98–125 (Harper and Row, New York, 1986).
19. Sibley, C. G. & Ahlquist, J. E. *Phylogeny and Classification of Birds* (Yale Univ. Press, 1990).
20. Martins, E. P. & Garland, T. Jr. *Evolution* **45**, 534–557 (1991).
21. Anderson, S. et al. *Nature* **290**, 457–465 (1981).
22. Kawasaki, E. in *PCR Protocols: A Guide to Methods and Applications* (eds Innis, M. H., Gelfand, D. H., Sninsky, J. J. & White, T. J.) 146–152 (Academic, San Diego, 1990).

ACKNOWLEDGEMENTS. We thank R. Schmidt and J. Felsenstein for advice and discussion; M. I. Ullah and A. R. Wani, the chief wildlife wardens of Kashmir, for help with field logistics; K. Marchetti and N. Jamdar for field assistance; D. Butcher for discussion and programming assistance; P. Berthold, H. Higuchi, J. Tiainen and C. Wood for field collections; and T. Case, H. L. Gibbs, P. Harvey, T. Langen, L. Liou, J. Losos, D. Schluter and P. Smouse for comments on the manuscript. This work was funded in part by grants from the Sigma Xi Society to A.D.R.

Travel at low energetic cost by swimming and wave-riding bottlenose dolphins

T. M. Williams, W. A. Friedl, M. L. Fong, R. M. Yamada, P. Sedivy & J. E. Haun

Naval Oceans Systems Center Hawaii Laboratory, PO Box 997, Code 511, Kailua, Hawaii 96734, USA

OVER the past 50 years there has been much speculation about the energetic cost of swimming and wave-riding by dolphins^{1–11}. When aligned properly in front of the bow of moving ships^{1–3}, in the stern wake of small boats^{4,5}, on wind waves⁶, and even in the wake of larger cetaceans^{7–9}, the animals appear to move effortlessly through the water without the benefit of propulsive strokes by the flukes. Theoretically, body streamlining as well as other anatomical and behavioural adaptations contribute to low transport costs in these animals. The economy of movement permitted by wave-riding has been perceived as an energetic advantage for the swimming dolphin^{2,3,10}, but has been hard to prove in the absence of physiological data for exercising cetaceans. Here we determine the aerobic and anaerobic costs of swimming and wave-riding in bottlenose dolphins and find that the minimum cost of transport for swimming dolphins is $1.29 \pm 0.05 \text{ J kg}^{-1} \text{ m}^{-1}$ at a cruising speed of 2.1 m s^{-1} . Aerobic costs are nearly twice as high for swimming seals and sea lions, and 8–12 times higher for human swimmers¹². Wave-riding by dolphins provides additional benefits in terms of speed. The results indicate that behavioural, physiological and morphological factors make swimming an economical form of high-speed travel for dolphins.

To determine whether there is an energetic saving associated with wave-riding, we trained bottlenose dolphins to follow a moving boat (Fig. 1). Freely swimming dolphins remained at least 1 m below the water surface and were positioned outside the boat's wake. Wave-riding dolphins preferred positions within the stern wake at a depth of $\sim 0.5 \text{ m}$. Heart rate, respiratory rate and post-exercise blood lactate concentration were measured for both positions.

The results show that the physiological responses of swimming dolphins differed from those of wave-riding animals. At swimming speeds up to 2.9 m s^{-1} , the physiological responses of dolphins followed patterns similar to those of other marine mammals^{12,13}: there was a graded, but not necessarily linear, increase in average heart rate, respiration rate and metabolic rate with increased swimming speed (Fig. 2; Table 1). Dolphins swimming at 2.1 m s^{-1} had respiratory rates, heart rates and blood lactate concentrations not significantly different from resting values (two-sample *t*-test¹⁴ at $P < 0.05$). We attribute these results to the relatively low energetic demands of swimming at this speed¹⁵ and to a pre-exercise, anticipatory response that occurred during the resting measurements. All three physio-

TABLE 1 Predicted metabolic rate and aerobic transport costs for swimming and wave-riding bottlenose dolphins

Speed (m s^{-1})	Average heart rate (beat min^{-1})	Metabolic rate ($\text{ml O}_2 \text{ kg}^{-1} \text{ min}^{-1}$)	Transport cost ($\text{J kg}^{-1} \text{ m}^{-1}$)
2.1	76.0 ± 1.0	8.07 ± 0.33	1.29 ± 0.05
2.9	126.3 ± 7.3	24.67 ± 2.41	2.85 ± 0.28
3.8	101.7 ± 6.6	16.55 ± 2.18	1.46 ± 0.19
(wave-ride)			

Only the highest speed was used for wave-riding. Metabolic rates were calculated from average heart rate using equation (1). Transport costs were calculated by dividing metabolic rate by swimming or wave-riding speed¹⁸. For these calculations, metabolic rate ($\text{ml O}_2 \text{ kg}^{-1} \text{ min}^{-1}$) was converted to metabolic energy (J^{23} assuming a caloric equivalent of $4.8 \text{ kcal per l O}_2$ and a conversion factor of $4.187 \times 10^3 \text{ J kcal}^{-1}$). Results are presented as the mean value $\pm 1 \text{ s.d.}$; $n = 4$ for each speed.

logical parameters were raised when swimming speed was increased to 2.9 m s^{-1} . For example, at 2.9 m s^{-1} , respiratory rate was 70% higher than resting values, blood lactate was 30% higher, and average heart rate was 62% higher. As the boat speed was increased to 3.8 m s^{-1} , the dolphins preferentially rode the stern wave. During wave-riding at this speed the average heart rate decreased, and respiratory rate and blood lactate concentration were similar to values obtained during swimming at 2.9 m s^{-1} .

By comparing the physiological responses of swimming and wave-riding dolphins travelling at the same speed, we can evaluate the relative effort associated with each mode of travel. At 3.8 m s^{-1} the respiratory rate of wave-riding dolphins was $5.5 (\pm 0.5 \text{ s.d.})$ breaths per min ($n = 3$); this compares with $8.8 (\pm 1.6 \text{ s.d.})$ breaths per min ($n = 5$) for freely swimming dolphins moving at the same speed. Lactate concentration for a swimming dolphin wearing the heart-rate harness was three times higher (103.6 mg dl^{-1} ; 11.6 mM) than the value for wave-riding dolphins moving at a similar speed. Although limitations in the instrumentation prevented the measurement of heart rate during high speed swimming, the results for respiratory rate and blood lactate concentration suggest that the energetic demands of wave-riding are considerably less than those of swimming at comparable speeds.

The aerobic transport costs of swimming and wave-riding dolphins also indicate an energetic advantage associated with wave-riding (Table 1). At 2.1 m s^{-1} , a routine cruising speed for wild bottlenose dolphins⁷, the cost of transport for swimming was $1.29 \pm 0.05 \text{ J kg}^{-1} \text{ m}^{-1}$. Transport costs doubled as swimming speed was increased to 2.9 m s^{-1} . With an increase in boat speed to 3.8 m s^{-1} , the dolphins switched to wave-riding. The resulting transport costs were lower than anticipated, and only 13% higher than for swimming at 2.1 m s^{-1} . Thus, the wave-riding dolphin obtains a significant advantage in speed for little energetic investment.

The hydromechanics that allow this energy savings during wave-riding are difficult to discern. Scholander¹ suggested that propulsion could be gained by simply 'leaning' a streamlined fluke into the forward slope of the bow wave. Others have detailed the theoretical balance between drag and buoyancy of the wave-riding dolphin, and lift forces created by the pressure field in front of a moving ship's bow^{6,9,10}. With sufficient assistance from the wave, propulsive stroking by the dolphin is reduced to the effort required to position the animal. Indeed, the wave-riding dolphin appears motionless once in position, with fluke movements limited to low-amplitude corrective strokes for repositioning. Only brief bursts of inhalation and exhalation at the water surface will interrupt the ride. Such mechanical advantages in the form of assisted propulsion by

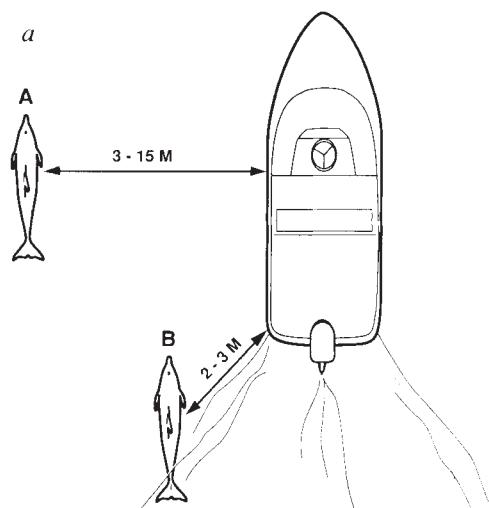
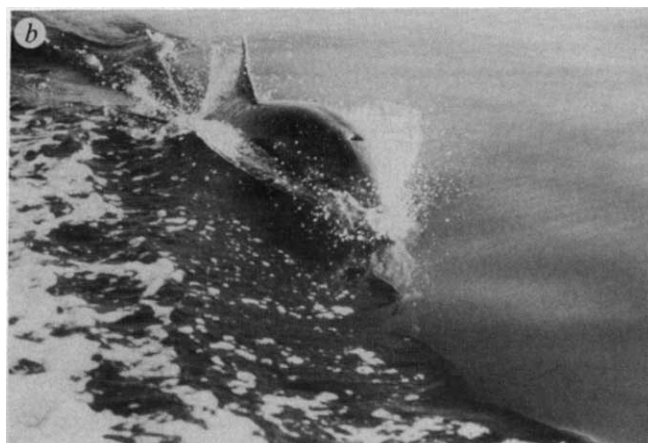


FIG. 1 *a*, Swimming position of bottlenose dolphins relative to the motorboat. Swimming animals (*a*) were positioned abeam outside the boat wake. The position for wave-riding (*b*) was approximated by the trainer and refined by the dolphins. *b*, Position of a bottlenose dolphin riding the stern wake of a boat. Wave-riding dolphins remained ~ 0.5 m below the water surface except for brief respiratory events (pictured). (NOSC photo.)

METHODS. We trained two adult Atlantic bottlenose dolphins (*Tursiops truncatus*; mean body weight, 145 kg) to match their swimming speed with that of a motorboat (Boston Whaler, Model 21 ft). Position of the animals relative to the boat was determined by a trainer and maintained with signals presented through a submerged hydrophone. The swimming dolphin was maintained in positions that avoided turbulent areas around the boat. Both dolphins preferentially swam about 1 m ($>2.5 \times$ body diameter) below the water surface during the swimming tests. At this depth the augmentation



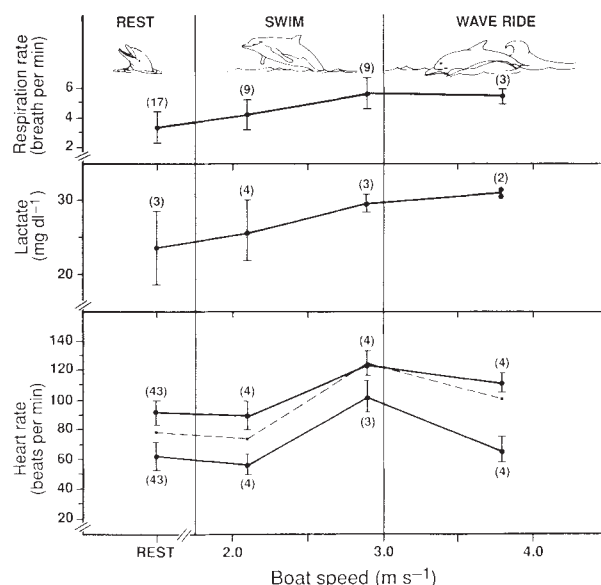
in drag from surface waves is reduced¹⁰. Surface intervals were limited to respiratory periods and accounted for $<10\%$ of the experimental period. During wave-riding each dolphin was allowed to refine its position within the stern wake. As a result, the distance between the boat and the dolphin varied with both position and speed. A 14-m-deep ship channel in Kaneohe Bay, Oahu, Hawaii was used for the experiments. Water currents are generally less than 0.3 m s^{-1} in this area and were considered negligible during the tests. Each session consisted of a 5-min pre-exercise rest period, and a 5-6-min warm-up swim, followed by a 10-25-min experimental swim at a single speed. Depending on the boat speed, most sessions covered ~ 1 -2 nautical miles (1.8-3.7 km) in a straight-line course. Speed of the boat was monitored with a Speed Log recorder calibrated before and after the experiments by timed trials over a measured course. Mean surface water temperature was $24.6 \pm 1.2^\circ\text{C}$.

FIG. 2 Heart rate, blood lactate and respiration rate for resting and active dolphins. Mean values ± 1 s.d. are presented for each parameter. Numbers in parentheses denote the number of experimental sessions. Heart rate of the dolphins cycled between periods of tachycardia following a breath and bradycardia during submergence. Upper and lower solid lines for heart rate show levels of bradycardia and tachycardia, respectively. The dashed line is the average heart rate at each speed.

METHODS. Electrocardiograph signals were continuously recorded (Birtcher, Model 365) from two 8.5-cm suction cup electrodes on the skin; a nylon harness kept the cups in place during swimming. The harness was small ($<5\%$ of the frontal area of the dolphin) and positioned close to the theoretical transition point for boundary layer flow²⁴. In view of this, its effect on energetic cost at cruising speeds was considered insignificant. Note that the effect of the harness will increase with speed; actual maximum speeds of the unencumbered dolphin may be higher than determined here. A plate electrode (3.0-cm diameter) beneath each suction cup was positioned on the sternum between the pectoral fins and mid-lateral in the axillary area of the right pectoral fin. Wires from the electrodes were braided and trailed to the side of the swimming animal. Respiratory rate was measured for 1 min at 5-min intervals by observers with stopwatches. Metabolic rate (V_{MR}) was calculated from average heart rate ($\bar{V}_{\text{HR}}$) using the equation

$$V_{\text{MR}} (\text{ml O}_2 \text{ kg}^{-1} \text{ min}^{-1}) = 0.33 \bar{V}_{\text{HR}} (\text{beats per min}) - 17.01 \quad (1)$$

(correlation coefficient, $r=0.98$) determined for the two dolphins pushing against a load cell^{25,26}. During the load cell tests heart rate was measured with two cross-chest electrodes; metabolic rate was determined simultaneously by open flow respirometry. The equation is based on $n=106$ steady-state sessions spanning loads of 0-185 kg, and average heart rates from rest (78.6 ± 7.2 beats per min) to maximum levels (139.4 ± 4.1 beats per min). Likewise, metabolic rates spanned resting to maximum values (T.M.W. *et al.*, manuscript in preparation). Average heart rate of the swimming dolphins used for metabolic calculations fell within this range of values. Samples for lactate analyses were obtained from blood vessels located in



the dolphin's fluke. Collection was facilitated by training the dolphins to present their flukes voluntarily for sampling. Using this technique, we obtained all samples within 2 min of completion of exercise. Chilled samples were centrifuged and the serum analysed for lactate concentration (YSI Industrial Analyzer, Model 27). The analyser was calibrated daily with standards spanning the experimental range of values.

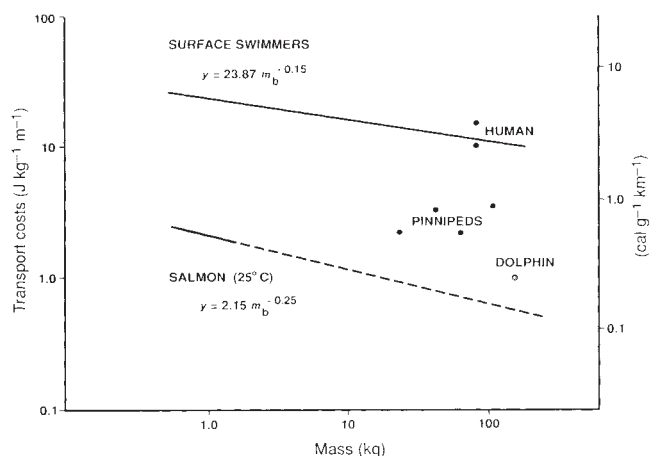


FIG. 3 Cost of transport for swimmers. The results for humans²⁰, pinnipeds^{12,19} and dolphins (this study) are shown. The upper solid line denotes the best-fit regression for surface swimmers and includes data for ducks, penguins, mink, muskrat, sea otters and humans²⁷; the lower line represents the best fit regression (solid) and extrapolation (dashed) for swimming salmon¹⁷. The extrapolation is based on that in ref. 18; y represents transport costs in $\text{J kg}^{-1} \text{m}^{-1}$; m_b is body mass in kg.

wave-riding may explain the phenomenal sustained swimming speeds of more than 8 m s^{-1} originally reported for dolphins at sea¹⁶.

Whether freely swimming or wave-riding, dolphins have low minimum transport costs compared with many swimmers (Fig. 3). The minimum cost of transport for swimming bottlenose dolphins was only 2.1 times the predicted value for a hypothetical dolphin-sized fish¹⁷. During wave-riding the costs increased slightly to 2.4 times predicted levels. It is remarkable that the difference in swimming costs between fishes and dolphins is not larger, especially if the energetic demands of endothermy and the additional drag associated with the heart-rate harness used here are considered. The hydrodynamic shape of the dolphin may play a part; transport costs determined in this study are between theoretical values based on laminar and turbulent water flow around the body¹⁸.

Comparisons with other mammalian swimmers also demonstrate the energetic advantages of dolphin swimming. Transport costs for swimming pinnipeds (seals, sea lions) are 2.5–5.7 times the predicted values for comparably sized fish^{12,19}. Humans are exceptionally inefficient surface-swimming mammals; regardless of the stroke used we maintain transport costs that are 8–12 times the value for dolphins and 21 times the predicted values for fish²⁰.

Dolphins and porpoises may use several behavioural strategies to reduce energetic costs during high-speed swimming^{21,22}. They preferentially seek out bow and stern wakes (this work, and refs 4, 5), and will match the speed of moving ships for minutes to hours¹⁰. We have found that what appears to be playful behaviour to the casual observer on ship also provides an economical (albeit not free) ride for the dolphin. □

Received 2 August 1991; accepted 6 January 1992.

- Schlander, P. F. *Science* **129**, 1085–1087 (1959).
- Norris, K. S. & Prescott, J. H. *Univ. Calif. Publ. Zool.* **63**, 291–402 (1961).
- Shane, S. H., Wells, R. S. & Wursig, B. *Mar. Mammal Sci.* **2**, 34–63 (1986).
- Ridgway, S. H., Scronce, B. L. & Kanwisher, J. *Science* **166**, 1651–1654 (1969).
- Ridgway, S. H. in *Mammals of the Sea: Biology and Medicine* (ed. Ridgway, S. H.) 590–747 (Thomas, Springfield, Illinois, 1972).
- Woodcock, A. H. & McBride, A. F. *J. exp. Biol.* **28**, 215–217 (1951).
- Wursig, B. & Wursig, M. *Fishery Bull.* **77**, 399–412 (1979).
- McBride, A. F. & Kritzer, H. *J. Mammal.* **32**, 251–266 (1951).
- Lang, T. G. in *Whales, Dolphins and Porpoises* (ed. Norris, K. S.) 410–432 (Univ. Calif. Press, Berkeley, 1966).
- Hertel, H. in *The Biology of Marine Mammals* (ed. Andersen, H. T.) 31–63 (Academic, New York, 1969).
- Gray, J. *J. exp. Biol.* **13**, 192–199 (1936).
- Williams, T. M., Kooyman, G. L. & Croll, D. A. *J. comp. Physiol. B* **160**, 637–644 (1991).
- Pongonis, P. J., Kooyman, G. L., Zornow, M. H., Castellini, M. A. & Croll, D. A. *J. comp. Physiol. B* **160**, 473–482 (1990).
- Zar, J. H. *Biostatistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1974).
- Williams, T. M. & Kooyman, G. L. *Physiol. Zool.* **58**, 576–589 (1985).
- Lang, T. G. & Pryor, K. *Science* **152**, 531–533 (1966).
- Brett, J. R. *J. Fish. Res. Board Can.* **21**, 1183–1226 (1964).
- Schmidt-Nielsen, K. *Science* **177**, 222–228 (1972).

- Fedak, M. A. in *Diving in Mammals and Man* (eds Brubakk, A., Kanwisher, J. W. & Sundnes, G.) 11–32 (Tapir, Trondheim, Norway, 1986).
- Homer, I. *J. appl. Physiol.* **33**, 502–509 (1972).
- Au, D. & Weihs, D. *Nature* **284**, 548–550 (1980).
- Blake, R. W. *J. mar. Biol. Assoc.* **63**, 61–70 (1983).
- Bartholomew, G. A. in *Animal Physiology: Principles and Adaptations* (ed. Gordon, M. S.) 57–110 (Macmillan, New York, 1977).
- Aleyev, Yu. G. *Nekton* (Junk, The Hague, Holland 1977).
- Williams, T. M. et al. *J. mar. Biol. Assoc.* **71**, 727–728 (1991).
- Goforth, H. W. Jr. in *The Bottlenose Dolphin* (eds Leatherwood, S. & Reeves, R. R.) 559–574 (Academic, San Diego, 1990).
- Williams, T. M. *J. comp. Physiol. A* **164**, 815–824 (1989).

ACKNOWLEDGEMENTS. This work was supported by an American Society for Engineering Education fellowship to T.M.W. and includes procedures and results from an Office of Naval Research Project. All experimental procedures were evaluated and approved according to animal welfare regulations specified under NIH guidelines. We thank R. Davis, F. Fish and D. A. Pabst for discussion; J. P. Schroeder for organizing veterinary support and for reviewing the manuscript; P. E. Nachtigall, E. Huber, K. Keller, M. Magee and E. Rawitz for help with lactate measurements; and L. Kahikina for shipboard assistance.

Overcompensation and population cycles in an ungulate

B. T. Grenfell, O. F. Price, S. D. Albon* & T. H. Clutton-Brock

Zoology Department, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

ALTHOUGH theoretical studies show that overcompensatory density-dependent mechanisms can potentially generate regular or chaotic fluctuations in animal numbers, the majority of realistic single-species models of invertebrate populations are not overcompensatory enough to cause sustained population cycles^{1–3}. The possibility that overcompensation may generate cycles or chaos in vertebrate populations has seldom been considered. Here we show that highly overcompensating density-dependent mortality can generate recurrent population crashes consistent with those observed in a naturally limited population of Soay sheep. The observed interval of three or more years between crashes points to sharp ‘focusing’ of mortality over a narrow range of population density.

Although comparative analyses of density dependence in large mammal populations has yielded important general insights about the significance of population limitation sharply concentrated at high population levels^{4–6}, more refined analyses of particular systems have been hampered by a lack of detailed long-term demographic data, particularly concerning adult survival^{6–8}.

The Soay sheep population on St Kilda has been monitored in detail between 1959 and 1968 (ref. 9), and from 1985 to the

* Present address: Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK.

present¹⁰. Over these periods, the population has shown a repeated pattern of crashes (Fig. 1a). Our analysis of these patterns is based on the last six years, which are the most well documented demographically—over 95% of lambs are tagged, and their subsequent survival and reproductive output monitored¹¹. The strongest density-dependent effect is on annual survival¹¹, which declines sharply during the population crash, when up to 70% of all sheep die. This effect, which is particularly marked in lambs and males (Fig. 1b–d), is caused primarily by starvation when high sheep population densities deplete the plant standing crop¹⁰. The resulting highly nonlinear relationship between annual survival (S) and total population size (N) indicates relatively high and constant survival at low sheep densities, with a sharp decline above a critical population threshold. To compare this pattern with the equivalent insect data (for which the effects of overcompensation have been most explored³), we use a standard survival function from that literature^{3,12}

$$S = d / (1 + (aN)^b) \quad (1)$$

Here the parameters d , a and b , respectively, control the level of density-independent mortality, the threshold population size ($1/a$) above which density-dependent mortality occurs, and the strength of density dependence (which is overcompensating for $b > 1$)^{3,12}. Fitting this model to the survival data (Table 1; Fig. 1b–d) yields values of b in the range 6.08 (adult females) to 16.31 (male lambs). This range of values indicates a degree of overcompensating density dependence which is much higher than in most insect populations, where the observed upper limit is $\sim b = 5$ ³. We can explore the resulting dynamic consequences using a deterministic age- and sex-structured population model^{13–15} (Fig. 2) reflecting the seasonal pattern of sheep reproduction^{10,11}. As lambs reproduce in their first year of life, and the survival curves of Fig. 1b–d are all qualitatively similar in shape, the dynamics of this complicated formulation can be captured crudely by adding across age classes to obtain the following simple model for the dynamics of the total population

$$N_{t+1} = \lambda d N_t / (1 + (aN_t)^b) \quad (2)$$

where a , b and d here are average survival parameters, and $\lambda = (f+1)$, where f is the average individual reproductive rate. This is a standard model for populations with discrete non-overlapping generations^{1,3,16,17}, and its local stability properties can be characterized by the relationship between b and the product λd (which represents reproductive rate allowing for density-independent mortality)³. Figure 2A plots this boundary superimposed on the average b and λd values for lambs yearlings and adults, in comparison with the equivalent figures from a range of insect studies³. It illustrates the very high degree of overcompensatory density dependence exhibited by the Soay population—even though the sheep reproductive rate is low compared with most insect populations, it is still sufficiently high to generate the possibility of unstable dynamics. To explore the system's stability in more detail, we return to the full age-structured model (which also allows for a small degree of density-dependent limitation on recruitment to the population¹¹; see Fig. 2). Figure 2B shows a typical simulation of the full model, which exhibits 2-point-limit cycle behaviour over a wide range of parameter space. This basic pattern of crashes in alternate years is also robust to the addition of noise.

Although this result is consistent with the observed instability of the system (Fig. 1a), it does not capture the timing of crashes, which tend to occur at 3–4-year intervals^{10,11}. The discrepancy arises because the simulation does not produce a sufficiently 'deep' crash (compare Figs 1a and 2B) and correspondingly longer recovery period. It indicates that the population is experiencing a more overcompensatory density-dependent mortality than is implied by equation (1). This is corroborated by

a more refined analysis of the survival data (Table 1; Fig. 1b–d), which provides evidence of a constant survival proportion at high population density in yearling and adult sheep. The occurrence of such a lower bound on survival at high population density may be modelled by modifying equation (1) to

$$S = d / (1 + (aN)^b) + c \quad (3)$$

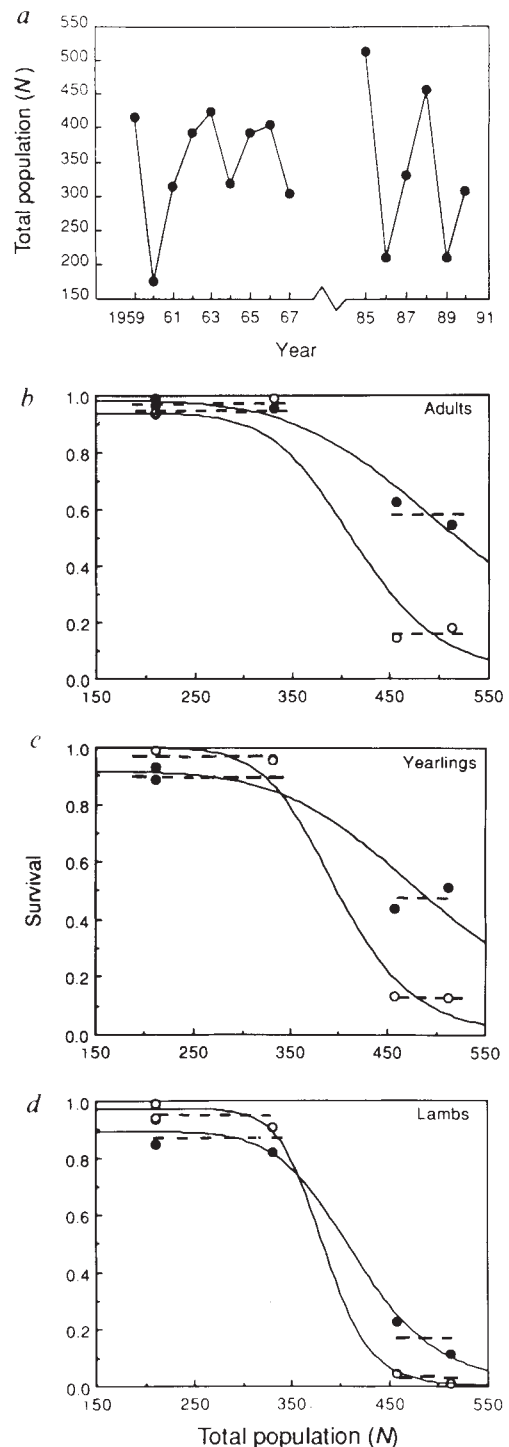


FIG. 1 a, Observed total population of sheep on Hirta, St Kilda in the periods 1959–67 (refs 9, 10) and 1985–90 (ref. 11). b–d, Annual survival proportion for male (○) and female (●) lambs (b), yearlings (c) and adults (d), in relation to total population size (N); based on data from Hirta, St Kilda, 1985–90. Fits of equation (1) (solid lines) and a simple model with constant survival rates below and above $N = 400$ (males, dashed lines; females, dotted lines) are shown (see Table 1 for details).

TABLE 1 Details of model fits to Soay sheep annual survival data by age and sex as a function of total population size (N)

	Survival model: $S = dN/(1 + (aN)^b)$				Survival model: $N < 400, S = S_1$ $N > 400, S = S_2$			Deviance difference (1)-(2)
	d	a	b	Deviance (1)	S_1	S_2	Deviance (2)	
Male lambs	0.971	0.00261	16.31	3.65	0.93	0.025	5.67	2.02
Female lambs	0.893	0.0024	9.89	1.88	0.86	0.168	6.8	4.92*
Male yearlings	0.99	0.00251	10.25	6.44	0.944	0.13	0.705	-5.74
Female yearlings	0.911	0.00201	6.21	10.0	0.922	0.472	1.746	-8.25
Male adults	0.937	0.0024	9.43	12.97	0.931	0.163	4.854	-8.12
Female adults	0.982	0.00192	6.08	9.5	0.963	0.581	11.44	1.94

Parameters were estimated by binomial maximum likelihood, from numbers surviving (with initial numbers as denominator). The basic survival model (equation (1)) is compared with a simple linear model in which survival takes different constant values below and above a population thresholds of $N=400$. Goodness-of-fit is assessed by the residual deviance of the fit²³, which is roughly chi squared with 3 degrees of freedom (equation (1)) and four degrees of freedom (linear threshold model). Differences between the deviance of the two models indicate that the threshold model is at least as good a fit as equation (1), except for lambs, where the nonlinear model is a significantly better fit (asterisk) for males. These statistical comparisons should be interpreted cautiously because the assumption of binomial errors ignores variation between years due, for example, to environmental variations (although these do not seem to be a strong influence on the observed dynamics¹¹).

where c is the lower survival limit. The implication of constant survival at the high end of observed population density (Fig. 3A) is a much greater average degree of density dependence ($b > 15$), which is concentrated in a relatively narrow population range ($350 < 1/a < 450$); but the extent of this effect (measured by b) cannot be estimated accurately from the data because survival proportions are either high or low (Fig. 3A). The implications of this focusing of mortality are explored in Fig. 3B-D, which shows the results of simulations based on a lower limit to yearling and adult survival. Figure 3B displays the dynamic implications of uncertainties in b in terms of a bifurcation diagram^{18,19}. For relatively low degrees of overcompensation ($b < 5$), the system is stable, then bifurcates into a 2-point cycle (analogous to Fig. 2B) as b increases ($8 < b < 15$). It then bifurcates further into a chaotic regime ($20 < b < 47$), settling eventually into a three-point cycle at high b ($b \geq 47$, when survival is

effectively a step function). This three-year cycle (Fig. 3C) corresponds more closely to the observed data, and Fourier analysis (Fig. 3D) shows this transition from 2-point to longer term cycles. Clearly, this comparison should be interpreted cautiously as it is based on only 6 years' observed data, but the correspondence of observed and simulated population totals (Fig. 3D) is encouraging. The observed summer total population of 440 individuals in 1991 is consistent with the model's prediction of another crash during the forthcoming winter (1991/92), and this will serve as another test of the model's assumptions.

This analysis used detailed demographic data to provide a proximate explanation for the spectacular recurrent crashes of Soay sheep populations on St Kilda in terms of highly overcompensating density-dependent mortality. The other important parameter here is the relatively high net reproductive rate of

FIG. 2 A, Local stability boundary for equilibria of the simple non-age structured model (equation (2))³ in terms of the degree of overcompensatory density dependence (b) and the average net reproductive rate (λd). The b estimates were calculated from binomial maximum likelihood fits to the total survival proportion in each age class (L, lambs; Y, yearlings; A, adults); the λd estimates are of average per capita fecundity in each age class (allowing for neonatal mortality), in the population range $N=300$ to 400 (the range of the model equilibrium). Note that this overestimates the stability of the model as overcompensation due to reproductive density dependence is ignored. The average of the sheep estimates is denoted by a filled circle and equivalent estimates for insect populations³ (open circles) are included for comparison. B, Numerical simulation of a more complex age-structured model, which is expressed in the following matrix equation

$$\begin{pmatrix} n_{f1} \\ n_{m1} \\ n_{f2} \\ n_{m2} \\ n_{f3} \\ n_{m3} \end{pmatrix}_{t+1} = \begin{pmatrix} f_1 & 0 & f_2 & 0 & f_3 & 0 \\ f_1 & 0 & f_2 & 0 & f_3 & 0 \\ S_{f1} & 0 & 0 & 0 & 0 & 0 \\ 0 & S_{m1} & 0 & 0 & 0 & 0 \\ 0 & 0 & S_{f2} & 0 & S_{f3} & 0 \\ 0 & 0 & 0 & S_{m2} & 0 & S_{m3} \end{pmatrix} \begin{pmatrix} n_{f1} \\ n_{m1} \\ n_{f2} \\ n_{m2} \\ n_{f3} \\ n_{m3} \end{pmatrix}_t$$

In this discrete-time formulation, the subscripts identify sex and age classes (1-3 for lambs, yearlings and adults, respectively), n are the numbers of sheep in different categories, with survival and female reproductive rates S and f respectively. Survival proportions are calculated from equation (1) (and Table 1), and reproductive rates (allowing for neonatal mortality) as the following linear density-dependent functions of total population size (N)¹¹, allowing for a 1:1 sex ratio at birth: adults $f=1.026-0.000966N$ ($R^2=0.511$); yearlings $f=1.271-0.00224N$ ($R^2=0.697$); lambs $f=1.0845-0.0022N$ ($R^2=0.961$).

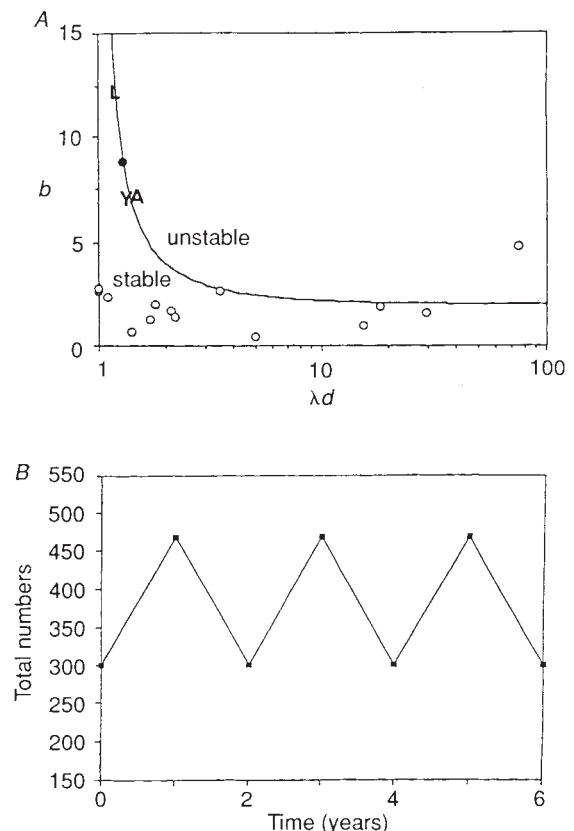
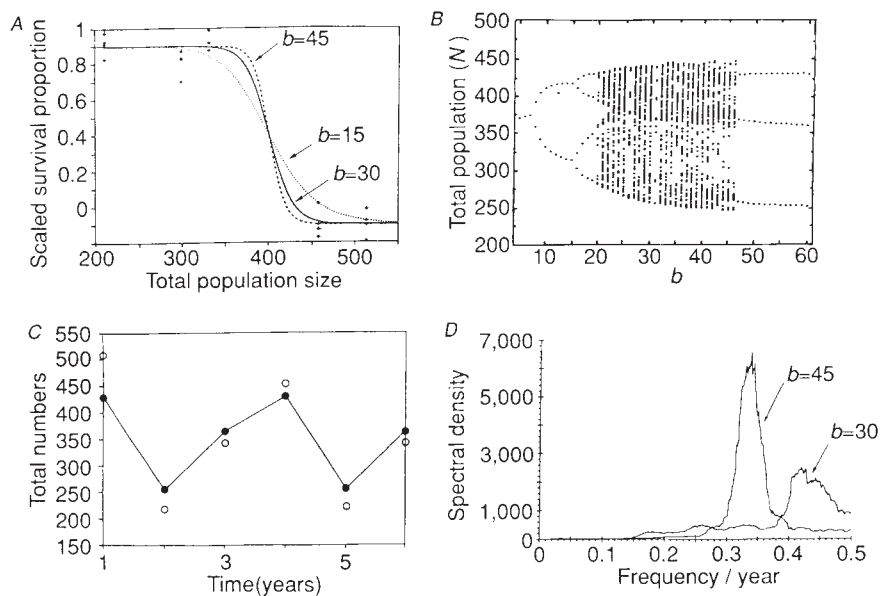


FIG. 3 A, Annual survival proportion of yearlings and adults, scaled relative to the upper and lower survival limits (S_1 and S_2) given in Table 1 (0 and 1 on this scale indicate survival proportions equal to the lower and upper limits respectively). The scaled survival curves are calculated from equation (3), with $d=1$, $c=0$ and $1/a=400$. The range of values of b indicate that a better fit to the data is achieved for $b > 30$, but that the data do not allow accurate estimation of a and b . B, Bifurcation diagram calculated from simulations of the age structured model (Fig. 2), 100 time points were sampled after a transient of 5,000 years. The model uses equations (1) (and Table 1) for lambs and equation (3) for yearlings and adults (in the latter case d and c are set to fit the limits in survival (S_1 and S_2) given in Table 1, and $1/a=400$). These results are conservative in that, if a lower survival limit is also assumed for lambs, cycles of longer period than 2 occur at lower values of b . C, Time plot of the model simulation at $b=50$, showing a 3-point cycle; for comparison, observed total population sizes for St Kilda from 1985–89 (1985 is year 1) are superimposed as points. D, Spectral analysis of simulations from B, showing the transition from 2-point to longer term cycles as b increases. Spectra were calculated from 1,000 time points after a transient of 5,000 years, and smoothed with a 25-point moving average. Adding random noise to the simulations does not alter these results qualitatively. In simple models (equation (2)), the population threshold for density dependence ($1/a$) does not affect the dynamics of the model³. Although in this more detailed



age-structured formulation $1/a$ does influence the period of oscillations in a complicated way, such variations do not affect the conclusion that a longer interval between crashes indicates higher degrees of overcompensatory density dependence.

the sheep population (Fig. 2A). Although a minority of ewes produce more than one lamb each year, the post-crash population recovery rate (and consequent instability) depends more on the ability of lambs to reproduce in their first year of life than on a high fecundity at later ages. The average observed period of 3–4 years between crashes, and the evidence of a lower observed limit to survival at high population density in older sheep, also indicate that there is a sharp peak in mortality over the population range of 350–450 individuals (Fig. 1a). We are now investigating the impact on this survival pattern of underlying factors, such as the basic plant–herbivore relationship and parasitism¹¹. We aim to clarify the relationship between Soay sheep population cycles and other well-documented vertebrate population oscillations^{20–22}.

The overcompensatory cycles in the Soay sheep population are markedly different from the case in many insect populations in the field, where a variety of more perfectly compensatory

density-dependent mechanisms reduce the intrinsic oscillatory tendency of predator–prey and host–parasite interactions¹⁶. This compensatory population regulation is also apparent in red deer, where density-dependent constraints on juvenile mortality and other demographic parameters prevent a significant overshoot of the population equilibrium^{7,8}. It is the absence of such additional constraints on the sheep population that permits their instability.

Although there is also some evidence of strongly overcompensating density-dependence in other ungulates^{4,6}, only long-term studies can clarify whether the associated dynamics are oscillatory. According to previous attempts to explain vertebrate population cycles using comparative allometric rules⁵, sheep (weight ~10 kg) should cycle with a period of over 10 years. The much shorter cycle period observed here further underlines the importance of detailed studies in clarifying the dynamics of individual populations. □

Received 24 September; accepted 27 November 1991.

- Hassell, M. P., Lawton, J. H. & May, R. M. *J. Anim. Ecol.* **45**, 471–486 (1976).
- Schaffer, W. M. & Kot, M. *TREE* **1**, 58–63 (1986).
- Bellows, T. S. *J. Anim. Ecol.* **50**, 139–156 (1981).
- Fowler, C. W. *Ecology* **62**, 602–610 (1981).
- Peterson, R. O., Page, R. E. & Dodge, K. M. *Science* **224**, 1350–1352 (1984).
- Fowler, C. W. in *Current Mammalogy* (ed. Genoways, H. H.) Vol. 1, 401–441 (Plenum, New York, 1987).
- Clutton-Brock, T. H., Major, M. & Guinness, F. E. *J. Anim. Ecol.* **54**, 831–846 (1985).
- Clutton-Brock, T. H. & Albon, S. D. *Red Deer in the Highlands* (Blackwell Scientific, Oxford, 1989).
- Boyd, J. M. & Jewell, P. A. in *Island Survivors* (eds Jewell, P. A., Milner, C. & Morton-Boyd, J.) 360–373 (Athlone, London, 1974).
- Grubb, P. J. in *Island Survivors* (eds Jewell, P. A., Milner, C. & Morton-Boyd, J.) 242–272 (Athlone, London, 1974).
- Clutton-Brock, T. H., Price, O. F., Albon, S. D. & Jewell, P. A. *J. Anim. Ecol.* **60**, 593–608 (1991).
- Maynard Smith, J. & Slatkin, M. *Ecology* **54**, 384–391 (1973).
- Leslie, P. H. *Biometrika* **33**, 183–212 (1945).
- Leslie, P. H. *Biometrika* **35**, 213–245 (1948).
- Clark, C. W. *J. math. Biol.* **3**, 381–391 (1976).

- Hassell, M. P. & Anderson, R. M. in *Ecological Concepts: the Contribution of Ecology to an Understanding of the Natural World* (ed. Cherrett, J. M.) 147–196 (Blackwell Scientific, Oxford, 1988).
- Varley, G. C., Gradwell, G. R. & Hassell, M. P. *Insect Population Ecology* (Blackwell Scientific, Oxford, 1973).
- Holden, A. V. *Chaos* (Princeton University Press, NJ, 1986).
- Schaffer, W. M. *Ecology* **66**, 93–106 (1985).
- Bulmer, M. G. *J. Anim. Ecol.* **43**, 701–718 (1974).
- Watson, A., Moss, R., Rothery, P. & Parr, R. *J. Anim. Ecol.* **53**, 639–662 (1984).
- Hudson, P. J. & Dobson, A. P. in *Red Grouse Population Processes* (eds Lantz, A. M. & Lawton, J. H.) 5–19 (RSPB and Brit. Ecol. Soc., London, 1990).
- Aitkin, M., Anderson, D., Francis, B. & Hindle, K. *Statistical Modelling GLIM* (Oxford Science, Oxford, 1989).

ACKNOWLEDGEMENTS. We thank the National Trust for Scotland and the Nature Conservancy Council for permission to work on St Kilda and for their help; the Army units stationed on St Kilda and Benbecula, and P. Schreiber and R. Collier for their assistance; and A. MacColl, M. Loneragan, J. Pemberton and D. Rubenstein for discussion. D. Bancroft, R. Clark, D. Green, F. Gulland, J. Kinsley, J. Pemberton, A. Robertson and I. Stevenson all made important contributions to this work, as did a substantial numbers of volunteers, including several members of the Mammal Conservation Trust. The work was supported by grants from NERC, SERC and the Mammal Conservation Trust (T.C.B., S.D.A. and O.F.P.), and the Wellcome Trust (B.T.G.).

P-type calcium channels blocked by the spider toxin ω -Aga-IVA

Isabelle M. Mintz*, Virginia J. Venema†, Kristine M. Swiderek‡, Terry D. Lee‡, Bruce P. Bean* & Michael E. Adams†§

* Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA

† Department of Entomology, University of California, Riverside, California 92521, USA

‡ Beckman Research Institute, Division of Immunology, Duarte, California 91010, USA

VOLTAGE-DEPENDENT calcium channels mediate calcium entry into neurons, which is crucial for many processes in the brain including synaptic transmission, dendritic spiking, gene expression and cell death¹⁻⁵. Many types of calcium channels exist in mammalian brains⁶⁻¹⁹, but high-affinity blockers are available for only two types, L-type channels (targeted by nimodipine and other dihydropyridine channel blockers²⁰⁻²²) and N-type channels (targeted by ω -conotoxin²²⁻²⁶). In a search for new channel blockers, we have identified a peptide toxin from funnel web spider venom, ω -Aga-IVA, which is a potent inhibitor of both calcium entry into rat brain synaptosomes and of 'P-type' calcium channels⁸ in rat Purkinje neurons. ω -Aga-IVA will facilitate characterization of brain calcium channels resistant to existing channel blockers and may assist in the design of neuroprotective drugs.

In the search for new calcium channel blockers, we focused on venom from the funnel web spider, *Agelenopsis aperta*, components of which inhibit presynaptic Ca^{2+} channels in insects²⁷⁻²⁹, Ca^{2+} current in sensory neurons^{30,31}, dendritic spiking in Purkinje neurons⁸, and rat brain Ca^{2+} channels expressed in *Xenopus* oocytes^{17,32}. Whole *A. aperta* venom contains a mixture of acylpolyamines³³⁻³⁶ and peptides^{28,33,35} (Fig. 1a). Because depolarization-induced Ca^{2+} entry into rat brain synaptosomes is largely resistant to both ω -conotoxin and dihydropyridine blockers, we used it as a bioassay for novel Ca^{2+} channel antagonists. Potent inhibitory activity occurred in the peak labelled 'IVA' in Fig. 1a, which was further purified (Fig. 1b, c) to a single peptide, referred to subsequently as ω -Aga-IVA.

Amino acid composition analyses of reduced, pyridylethylated ω -Aga-IVA produced the following ratios (with values from sequence analyses included in parentheses): Asx 2.2 (2), Glx 2.2 (2), Ser 1.1 (1), Gly 7.4 (8), Arg 4.1 (4), Thr 2.5 (2), Ala 2.1 (2), Pro 2.0 (2), Tyr 1.0 (1), Met 1.8 (2), Cys 6.1 (8), Ile 3.3 (4), Leu 3.0 (3), Lys 5.2 (6). Ultraviolet absorbance spectra showed the additional presence of tryptophan. N-terminal amino acid sequencing of ω -Aga-IVA yielded assignments for residues 1-42. Cleavage of the peptide with cyanogen bromide generated fragments 1-30, 31-42 and 43-48. Sequence analyses of the fragments showed that ω -Aga-IVA contains 48 amino acids including eight cysteines and one tryptophan (Fig. 1e), consistent with composition analyses and ultraviolet absorbance spectra of the intact, native peptide. The primary structure was confirmed by electrospray mass spectrometry, which yielded a relative molecular mass of $5,203 \pm 2$ (Fig. 1d), consistent with the sequence shown in Fig. 1e for a peptide containing four disulphide bonds ($5,202.3$ calculated for $\text{C}_{217}\text{H}_{360}\text{N}_{68}\text{O}_{60}\text{S}_{10}$). The spectrum of reduced, pyridylethylated ω -Aga-IVA gave a molecular mass of $6,052$ ($6,051.5$ calculated for $\text{C}_{217}\text{H}_{424}\text{N}_{76}\text{O}_{60}\text{S}_{10}$), confirming the presence of eight cysteine residues. Fast atom bombardment mass spectrometric analysis of the C terminal fragment (43-48) gave a value of 559.33 for the monoisotopic mass of the protonated molecule (559.64 calculated for $\text{C}_{24}\text{H}_{43}\text{N}_6\text{O}_9$) indicating that the C terminus is the free carboxylic acid.

ω -Aga-IVA is a potent blocker of Ca^{2+} channels in rat brain synaptosomes. The toxin produced $71 \pm 4\%$ ($n = 4$) block of voltage-dependent $^{45}\text{Ca}^{2+}$ entry at 200 nM , acting with 50% inhibition (IC_{50}) at about 20 nM (Fig. 2). Under the same conditions, the N-channel blocker ω -conotoxin GVIA (ref. 23, 24) had no effect at 500 nM . In contrast, when synaptosomes were prepared from chick brain, 100 nM ω -conotoxin GVIA eliminated $^{45}\text{Ca}^{2+}$ entry, but 500 nM ω -Aga-IVA had little effect (Fig. 2a), blocking only $19 \pm 5\%$ ($n = 13$). The contrasting actions of ω -Aga-IVA and ω -conotoxin, with reversed selectivities for

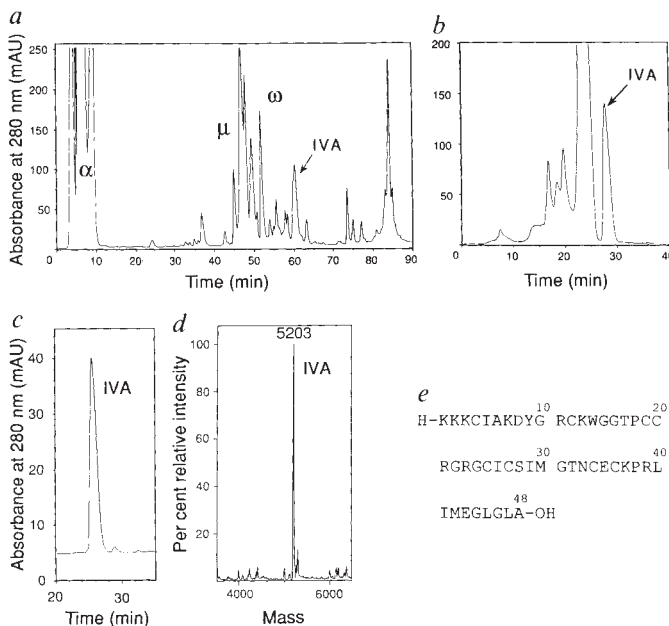


FIG. 1 Isolation and identification of ω -Aga-IVA from funnel web spider venom. **a**, Reversed-phase high performance liquid chromatography (RPLC) of whole venom ($20 \mu\text{l}$). The early eluting material (3–10 min) is a mixture of acylpolyamine α -agatoxins³³⁻³⁶; later eluting peaks are peptidic and comprise the μ - (35–50 min; Refs. 33, 35) and ω -agatoxins (48–65 min; Ref. 28). Under the solvent conditions used, ω -Aga-IVA (IVA) elutes as a component of the peak indicated at 60 min. **b**, A second RPLC step involving $300 \mu\text{l}$ equivalents of venom separates ω -Aga-IVA (eluting at 29 min) from several other peptides. **c**, Re-chromatography of this peak shows ω -Aga-IVA as a single peak eluting at 26 min. **d**, The electrospray mass spectrum of ω -Aga-IVA gave a series of multiply charged ions, which upon computer analysis yielded a value of $5,203 \pm 2$ for the peptide relative molecular mass, consistent with four disulphide bonds. **e**, Complete amino acid sequence of ω -Aga-IVA using single-letter amino acid code.

METHODS. Whole venom was fractionated on a Brownlee C_8 wide-pore analytical column ($4.6 \times 150 \text{ mm}$) with a linear gradient of aqueous acetonitrile in constant 0.1% trifluoroacetic acid (TFA) at 1.0 ml min^{-1} . The IVA peak was further purified using a Brownlee C_8 column and a linear gradient of n -propyl alcohol/water (0.33% per min) in constant 1% TFA. A third RPLC step used a Vydac wide-pore C_{18} analytical column and a linear gradient of aqueous acetonitrile in constant 0.1% TFA. Purification steps were determined using a bioassay for inhibition of $^{45}\text{Ca}^{2+}$ flux described in Fig. 2 legend. Methods for sequence analysis of the reduced pyridylethylated toxin, cyanogen bromide cleavage, and composition analyses for all amino acids except tryptophan were performed as described²⁸. Positive ion fast bombardment (FAB) spectra were obtained using an HX100HF (JEOL USA, Peabody, MA) double focusing magnetic sector instrument operated at 5-kV acceleration potential and a nominal resolution setting of $3,000$. Sample ions were desorbed with a 6 keV xenon atom beam. The sample matrix was 80% dithiothreitol: 20% dithioerythritol and 6 mM camphor sulphonic acid. Positive ion electrospray mass spectra were obtained using a TSQ-700 triple sector quadrupole (Finnigan MAT, San Jose, California). The electrospray needle was operated with a voltage differential of $3\text{--}4 \text{ kV}$. Samples were dissolved in 50% methanol and delivered into the atmospheric pressure electrospray ion source at $1 \mu\text{l min}^{-1}$. The multiply-charged electrospray spectrum was collected and analysed using Finnigan MAT's BIOMASS data reduction software. Mass values reported are for the average relative molecular mass of the neutral molecule.

§ To whom correspondence should be addressed.

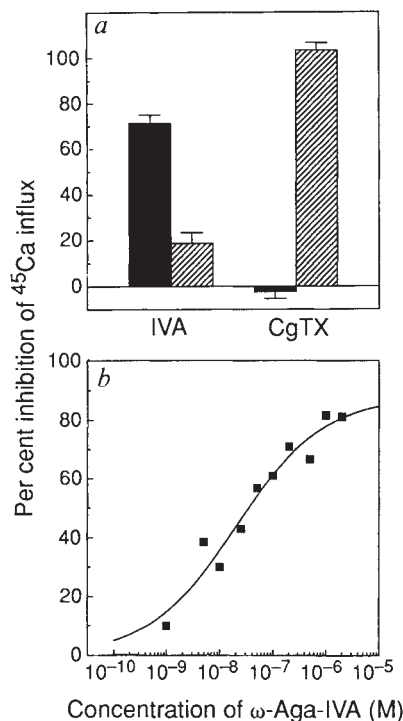
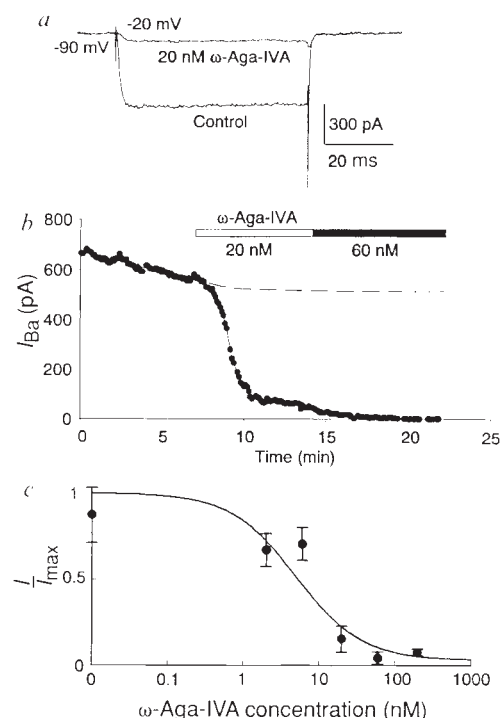


FIG. 2 ω -Aga-IVA and ω -conotoxin GVIA exhibit different selectivities for synaptosomal Ca^{2+} channels. **a**, Inhibition of potassium-stimulated $^{45}\text{Ca}^{2+}$ entry into synaptosomes from rat (solid bars) or chick (hatched bars) brain by ω -Aga-IVA (IVA) and ω -conotoxin GVIA (CgTX). **b**, Concentration dependence of ω -Aga-IVA inhibition of $^{45}\text{Ca}^{2+}$ flux in rat brain synaptosomes. **METHODS**. Synaptosomes were prepared from whole brains of 14–20-day-old Sprague-Dawley rats; all steps were performed at 4°C . Tissue was homogenized in isolation medium (320 mM sucrose, 5 mM TES, 5 mM EDTA, pH 7.4) using a Wheaton glass/teflon homogenizer and overhead stirrer. The homogenate was centrifuged for 10 min at $3,000g$ in a Sorval SA-600 rotor. The pellet was washed once, supernatants were pooled and centrifuged for 20 min at $12,500g$ (SA-600). Pellets were resuspended in fresh isolation medium and layered onto a discontinuous gradient of 5%, 9%, and 12% Ficoll 400 in isolation medium and centrifuged in a Beckman SW41-Ti rotor at $75,000g$ for 50 min. Synaptosomes were stored on ice as 1.5 mg pellets in 250 mM sucrose and 5 mM TES, pH 7.4 and used within 2 h. Synaptosomal protein was quantified by the biuret method⁴⁶ using BSA as a standard. For measurement of $^{45}\text{Ca}^{2+}$ influx, pellets were resuspended at 1 mg ml^{-1} in ice-cold basal saline containing (in mM) 145 choline chloride, 3.1 KCl, 1.2 MgSO_4 , 0.4 KH_2PO_4 , 0.5 CaCl_2 , 10 glucose, 15 HEPES pH 7.4, 1 mg ml^{-1} BSA, preincubated at 4°C for 20 min, then warmed to 30°C over 10 min. During 30 min preincubation, toxins were added at various times before depolarization. Synaptosomes were added as 125 μl aliquots to 50 μl of either basal or depolarizing (high K^+) saline containing $^{45}\text{Ca}^{2+}$ (0.05 μCi). Depolarizing saline was made by replacing choline chloride with potassium chloride to a final concentration of 45 mM KCl. Where 2 mM Co^{2+} was used as a control for complete block of $^{45}\text{Ca}^{2+}$ influx, synaptosomes were added to 50 μl of high K^+ saline containing CoCl_2 and $^{45}\text{Ca}^{2+}$. Influx was terminated after 5 s by addition of 600 μl wash solution, ice-cold basal saline without BSA or glucose and with 2 mM calcium. Synaptosomes were pelleted by centrifugation at $11,000g$ for 30 s using a Beckman Microfuge II. The supernatant was aspirated and pellets were washed twice in 500 μl ice-cold wash solution, and solubilized with 50 μl 5% (wt/vol.) SDS before addition of 1 ml scintillant (Liquiscint) and counting.

FIG. 3 Inhibition by ω -Aga-IVA of high-threshold current in rat Purkinje neurons. **a**, Ca^{2+} channel current elicited by 60 ms depolarization from -90 mV to -20 mV in a Purkinje neuron, recorded before and 7 min after application of 20 nM ω -Aga-IVA. The current was entirely high threshold¹⁹. Leak and capacitive currents were subtracted using scaled current for a step from -90 to -105 mV . **b**, Time-course of inhibition of Ba^{2+} channel current (I_{Ba}) by 20 nM ω -Aga-IVA, with subsequent addition of 60 nM toxin, same cell as **a**. Currents were elicited by 60 ms steps from -90 to -20 mV delivered every 7 s. After a lag of about 1 min due to slow toxin mixing the current decline could fit an exponential of time constant 50 s. The dashed line shows the extrapolated current in the absence of toxin, assuming that run-down is proportional to current size; the rate of run-down (time constant 33 min in the absence of toxin) was determined from the segment before toxin application. Currents were corrected for leak current and for a 23 pA outward non- Ca^{2+} channel current that persisted after addition of 600 μM CdCl_2 at the end of recording. **c**, Concentration-dependence of ω -Aga-IVA block of high-threshold Ca^{2+} channel current in Purkinje neurons. Each point represents mean $\pm$ s.e.m. of 4–8 measurements. The extent of block was determined 5 min after toxin application using 60 ms pulses from -90 mV to a test potential of -30 to -10 mV , delivered every 7 s, and correcting for extrapolated run-down as in **b**. Measuring after 5 min underestimates the steady-state block by 2 or 6 nM toxin, but run-down made accurate estimates at longer times in some cells impractical. As a control for run-down correction, applications of toxin-free solution were made in 4 cells and the current records analysed in the same way (0 nM toxin). The solid curve is drawn according to $0.03 + 0.97(1 + [\text{toxin}]/5\text{ nM})$. **METHODS**. Purkinje neuron cell bodies from 5–21 day-old rat pups were prepared as previously described¹⁹. Whole-cell Ca^{2+} channel currents⁴⁷ were recorded with 1.5–7-M Ω pipettes containing an internal solution of (in mM) 108 CsCH_3SO_3 , 4 MgCl_2 , 9 EGTA, 9 HEPES pH 7.4, 4 MgATP , 14 creatine phosphate (Tris salt), 1 GTP (Tris salt), 50 U ml^{-1} creatine phosphokinase. After establishing the whole-cell configuration in Tyrode's solution¹⁸, cells were transferred to a minichamber (55 μl) perfused with an external solution of (in mM) 5 BaCl_2 , 160 tetraethylammonium chloride, 0.1 EGTA, 10 HEPES pH 7.4. ω -Aga-IVA was dissolved in the same solution, containing 1 mg ml^{-1} cytochrome C to saturate non-specific peptide binding by containers. Concentrated toxin solutions were pipetted directly into the



minichamber. Toxin solutions were stored at -10°C and could withstand 5–6 freeze and thaw cycles over 4–5 weeks with little loss of potency. Currents were filtered at 10 kHz and digitized every 100 μs . Data were not accepted unless voltage errors from series resistance remaining after partial compensation were less than 5 mV. All experiments were done at 21 – 24°C .

rat and chick brain synaptosomes, suggests that they target different types of Ca^{2+} channels. Consistent with this, ω -Aga-IVA (200 nM) had little effect on binding of ^{125}I - ω -conotoxin to synaptosomal membranes, inhibiting only $10 \pm 5\%$ in rat synaptosomes and $3 \pm 2\%$ in chick synaptosomes. Thus, ω -Aga-IVA selectively targets a population of Ca^{2+} channels in rat brain synaptosomes that are distinct from N-type channels.

ω -Aga-IVA potentially blocked high-threshold 'P-type' Ca^{2+} current in rat Purkinje neurons (Fig. 3) which is resistant to both ω -conotoxin and dihydropyridines^{18,19}. The half-blocking concentration was 2–10 nM, and saturation occurred with 60–200 nM toxin (Fig. 3c), which blocked $94 \pm 2\%$ of the current ($n = 18$). The toxin therefore binds with high affinity to P-type Ca^{2+} channels in Purkinje neurons.

We also tested the toxin on high-threshold Ca^{2+} current in rat dorsal root ganglion (DRG) neurons, which is composed of ω -conotoxin-sensitive N-type current ($\sim 50\%$), dihydropyridine-sensitive L-type current ($\sim 20\%$), and a current resistant to both blockers ($\sim 30\%$)^{18,37}. ω -Aga-IVA produced saturating effects at 200 nM (Fig. 4), but blocked only 0–24% of the current (mean $13 \pm 2\%$, $n = 4$). Maximal block by ω -Aga-IVA did not prevent block by additional application of ω -conotoxin and nitrendipine (Fig. 4a), showing that N-type and L-type currents remained. In some DRG neurons, a small current remained with all three blockers present. Thus, ω -Aga-IVA is highly selective for block of P-type Ca^{2+} channels in Purkinje neurons over the mixture

of channels in DRG neurons (Fig. 4b). It remains to be seen whether the fraction of current blocked by ω -Aga-IVA in some DRG neurons is carried by P-type channels.

The whole venom of *A. aperta* has previously been found to inhibit a range of Ca^{2+} currents, some of which can be blocked by a venom fraction, FTX^{8,32}. The peptide ω -Aga-IVA is clearly distinct from FTX fraction, which is non-peptide and of much lower relative molecular mass. Whether the two toxins overlap in their Ca^{2+} channel selectivities remains to be tested.

N-type Ca^{2+} channels mediate some synaptic transmission in mammalian brains, as ω -conotoxin blocks some synapses^{3,38}, binds to rat brain synaptosomes^{39,40}, and can partially inhibit synaptosomal Ca^{2+} entry under conditions slightly different from those we used^{41–44}. The P-type Ca^{2+} channel blocker we describe here is much more effective than ω -conotoxin at inhibiting Ca^{2+} entry into rat brain synaptosomes, suggesting that P-type channels may have synaptic functions throughout the brain, not only in Purkinje neurons. This suggestion is consistent with recent immunocytochemical evidence for widespread P-type channels⁴⁵. Thus ω -Aga-IVA will be useful for investigating the roles of P-type Ca^{2+} channels in transmission, excitability, gene expression, and cell death. Use of ω -Aga-IVA as a biochemical probe will facilitate development of a new class of Ca^{2+} -channel blocking drugs targeted at P-type Ca^{2+} channels. □

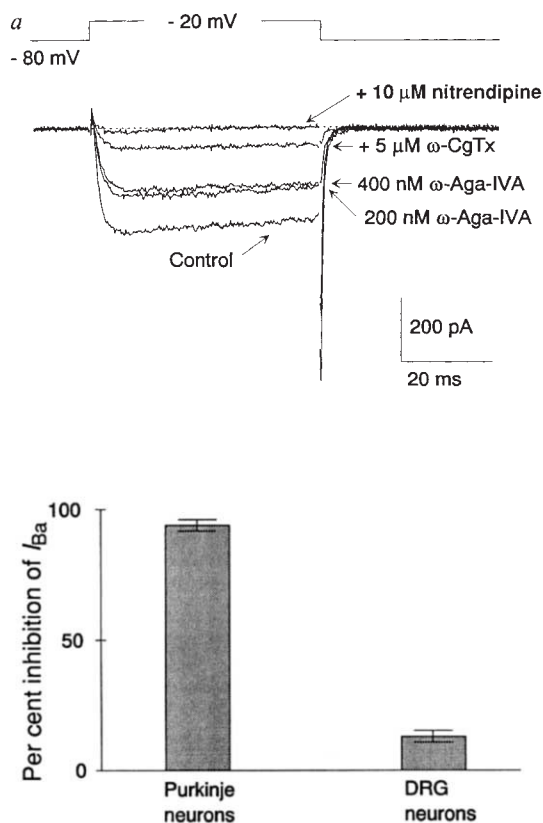


FIG. 4 a, Effect of ω -Aga-IVA in a rat dorsal root ganglion (DRG) neuron. This was the largest effect seen in application of 200 nM toxin to 4 cells. Depolarizations from -80 to -20 mV were given every 7 s in a freshly isolated DRG neuron¹⁸ studied with the solutions and protocols used for Purkinje neurons. Currents were recorded successively in control solution, in the presence of 200 nM ω -Aga-IVA, after an increase to 400 nM toxin, in the combined presence of 400 nM ω -Aga-IVA and $5 \mu\text{M}$ ω -conotoxin (ω -CgTx), and after addition of $10 \mu\text{M}$ nitrendipine to the other two blockers. b, Comparison of inhibition of high-threshold Ca^{2+} current by 200 nM ω -Aga-IVA in Purkinje neurons ($n=18$) and in DRG neurons ($n=4$).

Received 1 August; accepted 18 December 1991.

1. Llinas, R. & Yarom, Y. *J. Physiol., Lond.* **315**, 569–584 (1981).
2. Llinas, R. *Science* **242**, 1654–1664 (1988).
3. Kuan, Y. F. & Schofield, C. N. *Eur. J. Pharmacol.* **130**, 273–278 (1986).
4. Sheng, M. & Greenberg, M. E. *Neuron* **4**, 477–485 (1990).
5. Weiss, J. H., Hartley, D. M., Koh, J. & Choi, D. W. *Science* **247**, 1474–1477 (1990).
6. Yaari, Y., Hamon, B. & Lux, H. D. *Science* **235**, 680–682 (1987).
7. Takahashi, K., Wakamori, M. & Akaike, N. *Neurosci. Lett.* **104**, 229–234 (1989).
8. Llinas, R. R., Sugimori, M., Lin, J. W. & Cherksey, B. *Proc. natn. Acad. Sci. U.S.A.* **86**, 1689–1693 (1989).
9. Fisher, R. E., Gray, R. & Johnston, D. J. *J. Neurophysiol.* **64**, 91–104 (1990).
10. Bossu, J. L., Fagni, L. & Feltz, A. *Pflügers Arch.* **414**, 92–94 (1989).
11. O'Dell, T. J., Alger, B. E. *J. Physiol., Lond.* **436**, 739–767 (1991).
12. Snutch, T. P., Leonard, J. P., Gilbert, M. M., Lester, H. A. & Davidson, N. *Proc. natn. Acad. Sci. U.S.A.* **87**, 3391–3395 (1990).
13. Starr, T. V. B., Prystay, W. & Snutch, T. P. *Proc. natn. Acad. Sci. U.S.A.* **88**, 5621–5625 (1991).
14. Hui, A. *et al. Neuron* **7**, 35–44 (1991).
15. Leonard, J. P., Nargeot, J., Snutch, T. P., Davidson, N. & Lester, H. A. *J. Neurosci.* **7**, 875–881 (1987).
16. Mogul, D. J. & Fox, A. P. *J. Physiol., Lond.* **433**, 259–281 (1991).
17. Mori, Y. *et al. Nature* **350**, 398–402 (1991).
18. Regan, L. J., Sah, D. W. & Bean, B. P. *Neuron* **6**, 259–280 (1991).
19. Regan, L. J. *J. Neurosci.* **11**, 2259–2260 (1991).
20. Fox, A. P., Nowicky, M. C. & Tsien, R. W. *J. Physiol., Lond.* **394**, 149 (1987).
21. Fox, A. P., Nowicky, M. C. & Tsien, R. W. *J. Physiol., Lond.* **394**, 173 (1987).
22. Asakai, T. & Kasai, H. *Pflügers Arch.* **414**, 150 (1989).
23. Olivera, B. M., *et al. Science* **249**, 257–263 (1990).
24. Olivera, B. M., *et al. Science* **230**, 1338 (1985).
25. McCleskey, E. W., *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 4327–4331 (1987).
26. Plummer, M. R., Logothetis, D. E. & Hess, P. *Neuron* **2**, 1453 (1989).
27. Bindokas, V. P. & Adams, M. E. *J. Neurobiol.* **20**, 171–188 (1989).
28. Adams, M. E., Bindokas, V. P., Hasegawa, L. & Venema, V. J. *J. biol. chem.* **265**, 861–867 (1990).
29. Bindokas, V. P., Venema, V. J. & Adams, M. E. *J. Neurophysiol.* **66**, 590–601 (1991).
30. Scott, R. H., Dolphin, A. C., Bindokas, V. P. & Adams, M. E. *Molec. Pharmacol.* **38**, 711–718 (1990).
31. Mintz, I. M., Venema, V. J., Adams, M. E. & Bean, B. P. *Proc. natn. Acad. Sci. U.S.A.* **88**, 6628–6631 (1991).
32. Lin, J. W., Rudy, B. & Llinas, R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 4538–4542 (1990).
33. Adams, M. E., Herold, E. E. & Venema, V. J. *J. comp. Physiol.* **164**, 333–342 (1989).
34. Quistad, G. B. *et al. Biochem. biophys. Res. Commun.* **169**, 51–56 (1990).
35. Skinner, W. S. *et al. J. biol. Chem.* **264**, 2150–2155 (1989).
36. Jasys, V. J., *et al. J. Am. chem. Soc.* **112**, 6696–6704 (1990).
37. Scroggs, R. S. & Fox, A. P. *J. Neurosci.* **11**, 1334–1346 (1991).
38. Dutar, P., Rascol, O. & Lamour, Y. *Eur. J. Pharmacol.* **174**, 261–266 (1989).
39. Feigenbaum, P., Garcia, M. L. & Kaczorowski, G. *Biochem. biophys. Res. Commun.* **154**, 298–305 (1988).
40. Wagner, J. A., Snowman, A. M., Biswas, A., Olivera, B. M. & Snyder, S. H. *J. Neurosci.* **8**, 3354–3359 (1988).
41. Suszkiw, J. B., Murawsky, M. M. & Shi, M. *J. Neurochem.* **52**, 1260–1269 (1989).
42. Suszkiw, J. B., O'Leary, M. E., Murawsky, M. M. & Wang, T. *J. Neurosci.* **6**, 1349–1357 (1986).
43. Lundy, P. M., Frew, R., Fuller, T. W. & Hamilton, M. G. *Eur. J. Pharmacol.* **206**, 61–68 (1991).
44. Reynolds, I. J. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 8804–8807 (1986).
45. Hillman, D. *et al. Proc. natn. Acad. Sci. U.S.A.* **88**, 7076–7080 (1991).
46. Gornall, A. G., Bardawill, C. J. & David, M. M. *J. biol. Chem.* **177**, 751–766 (1949).
47. Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. *Pflügers Arch.* **391**, 85–100 (1981).

ACKNOWLEDGEMENTS. This work was supported by NIH and a fellowship from INSERM.

Volume-regulated chloride channels associated with the human multidrug-resistance P-glycoprotein

Miguel A. Valverde, Mario Díaz
& Francisco V. Sepúlveda

AFRC Institute of Animal Physiology and Genetics Research, Babraham,
Cambridge CB2 4AT, UK

Deborah R. Gill, Stephen C. Hyde
& Christopher F. Higgins

Imperial Cancer Research Laboratories, Institute of Molecular Medicine,
University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK

EXPRESSION of P-glycoprotein, the product of the *MDR1* gene, confers multidrug resistance on cell lines and human tumours (reviewed in refs 1, 2). P-glycoprotein (relative molecular mass 170,000) is an ATP-dependent, active transporter which pumps hydrophobic drugs out of cells³, but its normal physiological role is unknown. It is a member of the ABC (ATP-binding cassette) superfamily of transporters⁴, which includes many bacterial transport systems, the putative peptide transporter from the major histocompatibility locus, and the product of the cystic fibrosis gene (the cystic fibrosis transmembrane regulator, CFTR). CFTR is located in the apical membranes of many secretory epithelia⁵ and is associated with a cyclic AMP-regulated chloride channel^{6–8}. At least two other chloride channels are present in epithelial cells, regulated by cell volume and by intracellular Ca^{2+} , respectively^{9,10}. Because of the structural and sequence similarities between P-glycoprotein and CFTR^{4,11}, and because P-glycoprotein is abundant in many secretory epithelia^{12–14}, we examined whether P-glycoprotein might be associated with one or other of these channels. We report here that expression of P-glycoprotein generates volume-regulated, ATP-dependent, chloride-selective channels, with properties similar to channels characterized previously in epithelial cells.

Whole-cell chloride currents in NIH3T3 fibroblasts permanently transfected with the full-length human *MDR1* complementary DNA, were measured by patch-clamp techniques (Fig. 1). Only small currents were observed under isotonic control conditions (300 mOsm). By contrast, exposing cells to a bath solution of 220 mOsm induced sizeable inward and outward currents showing a clear decay at depolarizing pulses. This effect was readily reversed by returning the cells to isotonic medium (data not shown). Separate experiments ($n = 6$) demonstrated that intracellular hypertonicity also activated the channels. The increase in current was essentially complete one minute after changing to hypotonic medium, and little further change took place if left for up to 15 min in hypotonic conditions. The anionic nature of these currents was confirmed by replacing external chloride with gluconate, which largely decreased outward currents but had little effect on inward currents (Fig. 1a). The corresponding current-voltage relationships show a shift in the reversal potential from 0 to 40 mV, consistent with a $P_{\text{gluconate}}/P_{\text{Cl}}$ of 0.15. The current-voltage relationship for the current (Fig. 1a, lower panel) is outwardly rectifying. The experiments included ATP and GTP in the pipette solutions. When ATP and GTP were omitted from the pipette, no current could be activated by hypotonicity (not shown). The current could not be activated by treatment with a mixture of 10 μM 3-isobutyl-1-methylxanthine, 10 μM forskolin and 0.5 mM dibutyryl cAMP, known to activate the CFTR-dependent chloride channel^{6–8}. The characteristics of the volume-regulated currents measured here are similar to those reported previously in

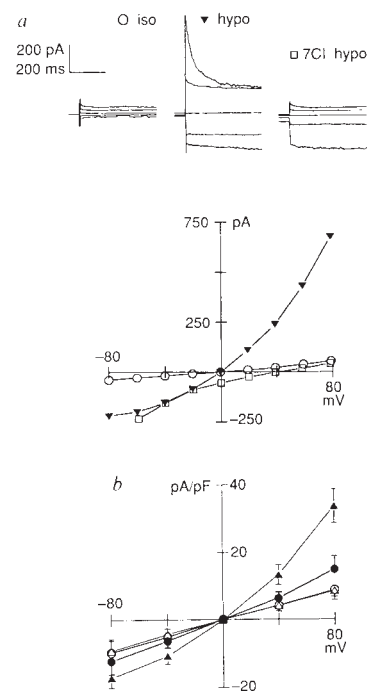
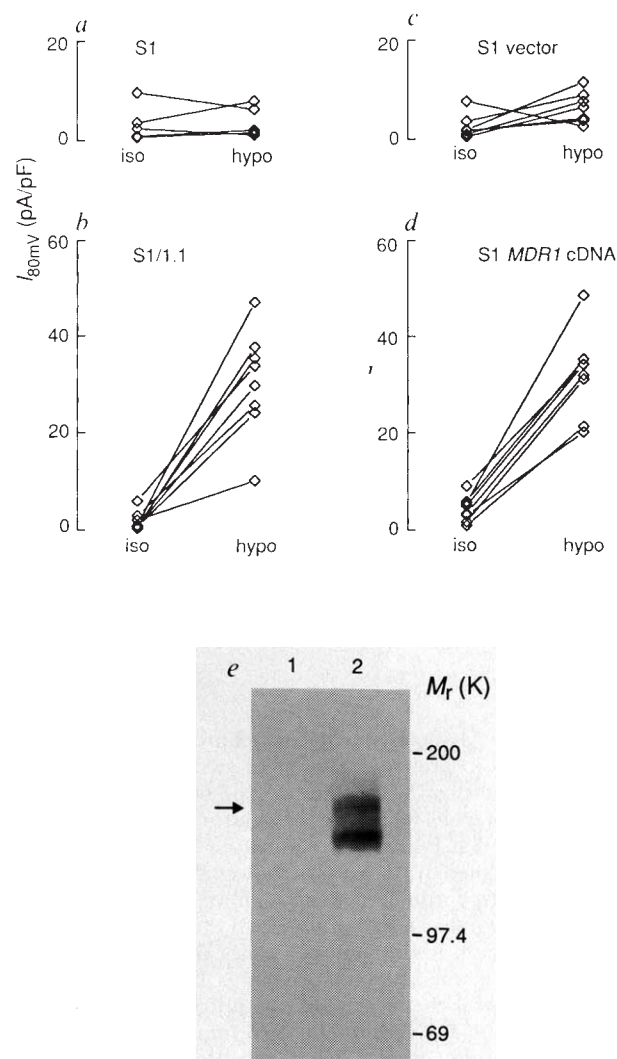


FIG. 1 Characterization of a chloride current in *MDR*-transfected NIH3T3 fibroblasts. **a**, Chloride currents measured for a single *MDR*-transfected NIH3T3 cell held at 0 mV and pulsed to voltages from -80 to +80 mV in 40 mV steps. The external bathing medium was changed from a normal isotonic solution (iso) to a hypotonic one (hypo), and then to a hypotonic bathing solution where all but 7 mM NaCl was replaced by sodium gluconate (7Cl⁻ hypo). Upper panel, typical recordings for a single cell; lower panel, corresponding current-voltage relations. Symbols as for the current traces in the upper part. **b**, Current-voltage relationships in NIH3T3 cells permanently transfected with human *MDR1* DNA, measured in isotonic medium (Δ) and 1 min after superfusion with hypotonic medium ($\blacktriangle$). The current-voltage relations in nontransfected, parental NIH3T3 cells were determined under the same conditions ($\circ$, $\bullet$, respectively). All current measurements were taken in the first 6 ms of the current pulse. The results shown are the means ($\pm$ s.e.m.) of 16 experiments for the transfected cell line, and 6 experiments for the nontransfected cell line. Similar experiments done on different days gave similar data. Currents measured in hypotonicity were significantly greater than those in isotonicity for transfected cells ($P = 0.0036$) but the difference in nontransfected cells did not reach significance ($P = 0.21$), as analysed by the *t*-test.

METHODS. Whole-cell currents were measured by the patch-clamp technique as described²¹. The ionic compositions of the pipette and bath solutions were chosen to measure chloride currents. The holding potential was 0 mV and currents were measured in response to 400-ms pulses from -80 to +80 mV in 20 or 40 mV steps. Voltage protocols, digitization and analysis were done using programs developed by J. Dempster (Strathclyde University). NIH3T3 cells are mouse fibroblasts; a derivative of these NIH3T3 cells (NIH3T3 *MDR1* c1) has been permanently transfected with human *MDR1* cDNA selecting for 1 $\mu\text{g ml}^{-1}$ colchicine^{22,23}. Western blots²⁴ confirmed that P-glycoprotein is undetectable in the parental NIH-3T3 line but is highly expressed in the NIH3T3 *MDR1* c1 cells. Cells were grown in 35-mm dishes in Dulbecco's modified Eagle's medium containing 10% fetal calf serum and used 1–2 days after subculturing. Solution changes were effected by directing a small jet of solution to the cell under study with a device similar to that described elsewhere²⁵. The pipette solution used in the experiments shown in **a** contained CsCl 140 mM, MgCl₂ 1.2 mM, EGTA 1 mM, ATP 2 mM, GTP 0.5 mM, HEPES 10 mM, pH 7.4; the corresponding isotonic bathing solution contained KCl 5 mM, NaCl 140 mM, CaCl₂ 1.3 mM, MgCl₂ 0.5 mM, HEPES 10 mM, pH 7.4; in the hypotonic solution NaCl was reduced by 35 mM; the low-Cl⁻ hypotonic solution (7Cl⁻ hypo) contained 105 mM sodium-gluconate instead of the NaCl. In **b** and in all other figures, the pipette solution was the same as in **a**, but with CsCl replaced by *N*-methyl-D-glucamine chloride (NMGCl), the isotonic bathing solution contained NMGCl 140 mM, CaCl₂ 1.3 mM, MgCl₂ 0.5 mM, HEPES 10 mM, pH 7.4, and the hypotonic solution had the NMGCl concentration reduced to 105 mM. The tonicity of the pipette and isotonic bathing medium was 300 mOsm, measured by freezing point depression. The osmolality of the hypotonic medium was 220 mOsm.

FIG. 2 Vaccinia-mediated, transient expression of P-glycoprotein and chloride currents in S1 cells. Cells were grown in Ham's F-10 medium supplemented with 10% fetal calf serum and used 48 hr after subculturing. Chloride currents were measured by whole-cell patch-clamping as described in the legend to Fig. 1. Data are presented for individual cells as peak currents following a +80 mV pulse under iso- and hypotonic conditions. *a*, Currents for S1 lung cells. *b*, Currents for S1/1.1 cells, a derivative of the S1 line which has been permanently transfected with human *MDR1* cDNA²⁶. Expression of P-glycoprotein in the S1/1.1 cells, but not the S1 cells, was confirmed by western blotting (data not shown). *c*, *d*, *e*, Chloride currents after transient expression of P-glycoprotein in S1 cells using a vaccinia system. Full-length human *MDR1* cDNA was obtained from P. Borst and the initiation codon modified by site-directed mutagenesis to generate an *NdeI* restriction site. The *MDR1* gene was sequenced after mutagenesis to ensure no additional changes had been introduced. The modified *MDR1* gene was cloned into plasmid pTM1 (B. Moss, personal communication) in two stages. First, the *NdeI*-*EcoRI* fragment of *MDR1*, and then the *EcoRI*-*DraI* fragment, were inserted between the *NcoI* and *StuI* sites in the pTM1 polylinker, such that *MDR1* was reconstructed under control of the phage T7 promoter. The resultant plasmid is designated pMDR7. The vaccinia-T7 hybrid expression system developed by Moss and coworkers²⁷ was used to express P-glycoprotein in S1 cells infected with vaccinia virus VTF7-3 encoding T7 RNA polymerase. Plasmid pMDR7, and separately the parent control plasmid pTM1, were transfected by a modification of the lipofectin procedure²⁸. This results in productive transfection and expression in 100% of cells²⁹. *c*, Chloride currents in S1 cells 12–15 h after vaccinia coinfection with the vector plasmid pTM1. *d*, Chloride currents in S1 cells 12–15 h after vaccinia coinfection with the *MDR1*-expression plasmid pMDR7. A *t*-test of the difference between hypo- and isotonicity gave *P*-values greater than 0.05 in *a* and *c* and smaller than 0.0003 in *b* and *d*. *e*, Western blot confirming vaccinia-induced transient expression of P-glycoprotein. Proteins were extracted by washing the cell suspensions twice in PBS and solubilizing the cell pellets with an equal volume of 1% (v/v) Triton X-100 (Sigma) in 10 mM Tris-HCl, pH 8.0 for 20 min on ice. Samples were centrifuged to remove cell debris and incubated at 37 °C for 30 min in Laemmli sample buffer³⁰. Proteins were separated by electrophoresis on a 7% SDS-polyacrylamide gel (bisacrylamide: acrylamide 1:30)³⁰, transferred to nitrocellulose membrane (Hybond-C super) in transfer buffer (25 mM Tris-Cl, pH 8.3, 192 mM glycine, 20% methanol, 0.04% SDS) and probed with the P-glycoprotein-specific polyclonal antibody 4077 (ref. 24). Immunodetection was by enhanced chemiluminescence (Amersham International). Lane 1, vector pTM1; lane 2, *MDR1* expression plasmid pMDR7. Relative molecular mass (*M_r*) markers are indicated in thousands (K). The upper band (arrow) is glycosylated P-glycoprotein; the lower band is presumably non-glycosylated and therefore incorrectly localized protein.



epithelial cells¹⁰, and distinct from those activated by cAMP and Ca^{2+} (ref. 9).

The increase in current on exposure to hypotonic medium was seen in all the transfected cells, whereas nontransfected cells showed little or no response (Fig. 1b). Similar results were obtained for a lung epithelial cell line (S1), and its derivative (S1/1.1) which has been permanently transfected with the *MDR1* gene (Fig. 2). S1 cells showed little or no hypotonicity-activated chloride current (Fig. 2a), whereas a large, volume-activated current was observed in all the S1/1.1 cells (Fig. 2b). The kinetics of this current were similar to those detected in the *MDR1*-transfected NIH3T3 cells.

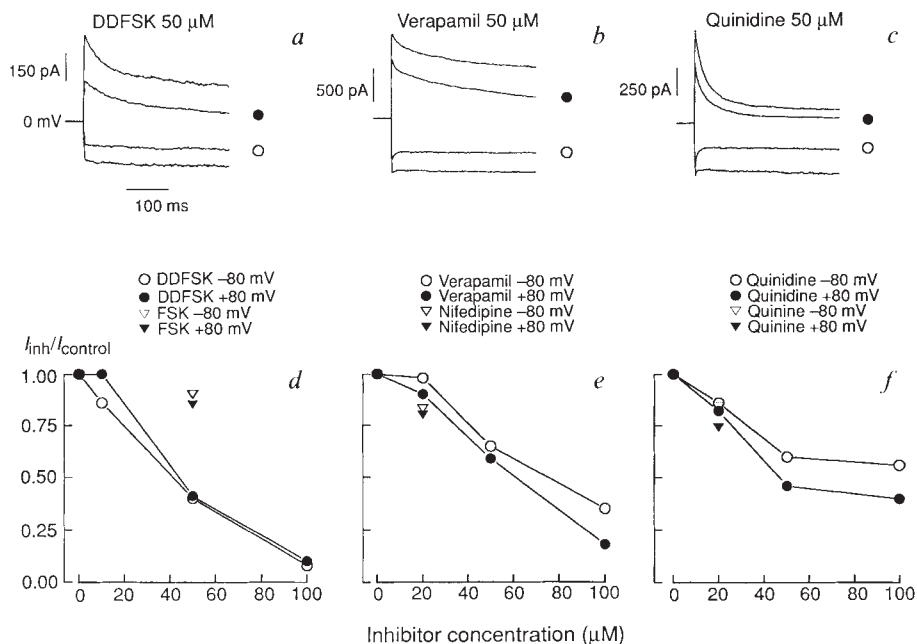
The volume-regulated chloride channels of *MDR1*-transfected cells were inhibited both by forskolin and by 1,9-dideoxyforskolin (Fig. 3). Both are inhibitors of P-glycoprotein function and interact with the protein¹⁵ although, unlike forskolin, 1,9-dideoxyforskolin does not interact with adenylate cyclase¹⁶, ruling out an effect through cAMP. The concentrations at which forskolin and dideoxyforskolin inhibited channel activity (Fig. 3) were similar to those at which they inhibit drug transport¹⁵. Other inhibitors of P-glycoprotein-mediated drug transport, including varapamil, quinine, nifedipine and quinidine, also inhibited the volume-activated chloride current (Fig. 3). The effects were fast (within 30 s of perfusion) and fully reversible. Conventional chloride-channel blockers, 5-nitro-2-(3-phenylpropylamino)benzoic acid (100 μM) and 4,4'-

diisothiocyanatostilbene-2,2'-disulphonic acid (100 μM), also inhibited these volume-regulated chloride currents (data not shown).

We did two other experiments to confirm an association between P-glycoprotein expression and the hypotonicity-activated chloride channels. First, *MDR1*-transfected NIH3T3 cells were grown in the presence of sense or antisense oligonucleotides against the 5' end of the *MDR1* gene (Fig. 4). The antisense oligonucleotide reduces both P-glycoprotein expression and drug accumulation¹⁷. The volume-activated chloride currents were unaffected by the sense oligonucleotide (Fig. 4a) but were abolished by the antisense oligonucleotide (Fig. 4b). A western blot demonstrated that the antisense oligonucleotide reduced P-glycoprotein levels by over 90%, compared with the sense oligonucleotide (Fig. 4c). Second, transient expression of P-glycoprotein from *MDR1* cDNA was directed in S1 cells using a vaccinia-based expression system (Fig. 2). Volume-activated chloride currents, with kinetic characteristics similar to those described above, were generated in cells transfected with the *MDR1*-expressing plasmid (Fig. 2d), but not in cells transfected with the parental plasmid vector (Fig. 2c). Western blotting confirmed P-glycoprotein expression after transfection (Fig. 2e).

Our data show that expression of P-glycoprotein correlates with the activity of ATP-dependent, cell volume-regulated chloride channels. Although P-glycoprotein functions as an active transporter to pump hydrophobic drugs from the cell, the

FIG. 3 Inhibition of chloride-channel activity by P-glycoprotein inhibitors. Volume-regulated chloride currents were activated and measured in NIH3T3 cells permanently transfected with the *MDR1* gene, as described in the legend to Fig. 1. *a, b, c*, Representative current traces taken in hypotonic medium in response to +80 or -80 mV pulses. Current traces were taken in the absence (no symbols) or presence of a given drug (filled symbols, +80 mV; open symbols, -80 mV). *d, e, f*, Effect of the drug is plotted as fractional inhibition of current against concentration. All effects were reversible and measurements given were taken 30 s after start of superfusion with the drug. DDFS, 1,9-dideoxyforskolin; FSK, forskolin.



chloride currents associated with P-glycoprotein expression are channel-like and are not consistent with ATP-dependent chloride transport, or with chloride co- or counter-transport with a drug molecule. The most straightforward interpretation of these data is that P-glycoprotein is itself a chloride channel, or a component thereof. We cannot, however, exclude the possibility that P-glycoprotein activates an endogenous chloride channel, although this would require that the channel pre-exists in an inactive form in different cell types. The hypothesis that P-glycoprotein is the channel protein is consistent with the finding that several chemically diverse drugs, known to inhibit drug transport by P-glycoprotein, also inhibit P-glycoprotein-dependent chloride currents. As P-glycoprotein has two ATP-binding domains, the ATP-requirement for chloride channel activity, presumably in a regulatory capacity, is not unexpected.

Three distinct chloride currents have been measured in epithelial cells, regulated by cAMP, Ca^{2+} and cell volume, respectively^{9,10}. The cAMP-regulated current has been associated with CFTR⁶⁻⁸. The P-glycoprotein-dependent chloride currents are distinct from those generated by CFTR expression, and have characteristics similar to those of volume-regulated chloride currents characterized previously in airway and other epithelial cells. These volume-activated conductances have been localized to the apical membranes of secretory epithelia^{18,19} and P-glycoprotein is also found in the apical membranes of secretory epithelia such as those lining small pancreatic ducts, proximal kidney tubules, columnar epithelia of colon and jejunum, the lumen of sweat ducts, and the endometrium¹²⁻¹⁴. Additionally, the regulatory volume decrease that occurs in some cells through activation of potassium and chloride channels is blocked by

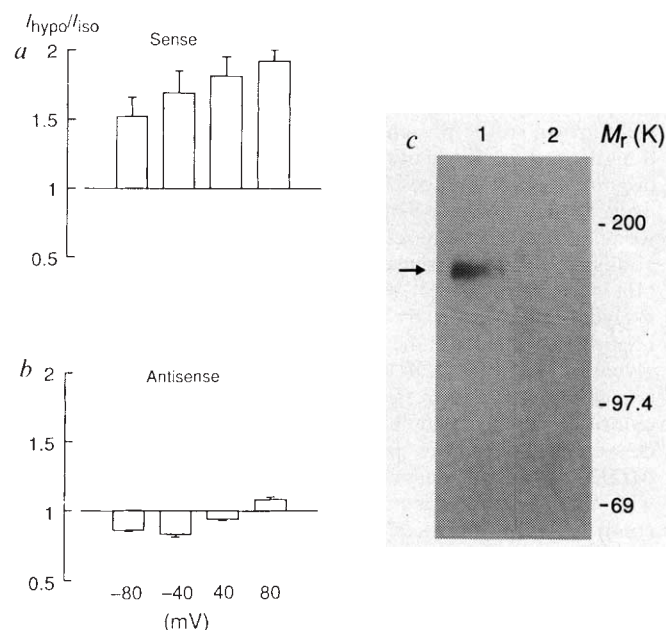


FIG. 4 Antisense inhibition of chloride channel activity and P-glycoprotein expression. NIH3T3 cells permanently transfected with the *MDR1* gene (see legend to Fig. 1) were grown, separately, in the presence of sense or antisense oligonucleotides corresponding to the 5' end of the human *MDR1* gene¹⁷. The oligonucleotides were included in the growth medium at $80 \mu g ml^{-1}$ for 4 days before patch-clamping; growth medium, including the oligonucleotides, was renewed every 24 h. Whole-cell currents were measured as described in the legend to Fig. 1*b*. Currents are expressed as a ratio of the current in hypotonic medium to that in isotonic medium as means ($\pm$ s.e.m.) of measurements made on six cells grown under each condition. *a*, Chloride currents for cells grown in the presence of the sense oligonucleotide. *b*, Currents for cells grown in the presence of the antisense oligonucleotide. The difference between hypo- and isotonicity was tested by paired *t*-test, giving *P*-values of 0.0027, 0.033, 0.0033 and 0.0029 for voltage pulses of -80, -40, 40 and 80 mV, respectively, in sense oligonucleotide-treated cells. The corresponding values for antisense-treated cells were 0.8, 0.15, 0.74 and 0.88. *c*, Western blot showing P-glycoprotein expression in cells grown with the sense and antisense oligonucleotides. Proteins were extracted from the same batches of cells used for patch-clamping, and electrophoresis and western blotting are described in the legend to Fig. 2. P-glycoprotein is indicated by an arrow. Equal amounts of total protein were loaded into each gel lane and confirmed after electrophoresis and blotting by staining. Lane 1, proteins from cells grown with the sense oligonucleotide; lane 2, proteins from cells grown with the antisense oligonucleotide. Densitometry using a Millipore Biolumager showed the antisense oligonucleotide reduced P-glycoprotein levels by over 90% compared with the sense oligonucleotide.

nifedipine and quinine²⁰; a similar effect of 1,9-dideoxyforskolin on regulatory volume decrease has been observed in isolated intestinal crypts (J. A. O'Brien, personal communication). Thus, it seems unlikely that the channels detected here are an artefact of P-glycoprotein overexpression.

The hypothesis that P-glycoprotein itself has chloride channel activity has several important implications besides those for multidrug resistance. First, P-glycoprotein may provide an

example of a single protein involved in both transport and channel functions; this has mechanistic implications. Second, as P-glycoprotein and CFTR are related to each other yet differ from other known channels⁴, they may define a new family of channel proteins. Third, if P-glycoprotein is associated both with transport and with channel activities, a function for CFTR as a transporter as well as a channel should also be considered. □

Received 2 October; accepted 18 December 1991.

- Gottesman, M. M. & Pastan, I. *Trends pharmac. Sci.* **9**, 54–58 (1988).
- Endicott, J. A. & Ling, V. A. *Rev. Biochem. Sci.* **58**, 137–171 (1989).
- Horio, M., Gottesman, M. M. & Pastan, I. *Proc. natn. Acad. Sci. U.S.A.* **85**, 3580–3584 (1988).
- Hyde, S. C. *et al. Nature* **346**, 362–365 (1990).
- Crawford, I. *et al. Proc. natn. Acad. Sci. U.S.A.* **88**, 9262–9266 (1991).
- Anderson, M. P. *et al. Science* **253**, 202–205 (1991).
- Kartner, N. *et al. Cell* **64**, 681–692 (1991).
- Tabcharani, J. A., Chang, X.-B., Riordan, J. R. & Hanrahan, J. W. *Nature* **352**, 628–631 (1991).
- Wagner, J. A. *et al. Nature* **349**, 793–796 (1991).
- Worrell, R. T., Butt, A. G., Cliff, W. H. & Frizzell, R. A. *Am. J. Physiol.* **256**, C1111–C1119 (1989).
- Riordan, J. R. *et al. Science* **245**, 1066–1073 (1989).
- Thiebaut, F. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 7735–7738 (1987).
- Cordon-Cardo, C. *et al. J. Histochem. Cytochem.* **38**, 1277–1287 (1990).
- Arceci, R. J., Croop, J. M., Horwitz, S. B. & Honsman, D. *Proc. natn. Acad. Sci. U.S.A.* **85**, 4350–4354 (1988).
- Morris, D. I., Robbins, J., Ruoho, A. E., Sutkowski, E. M. & Seaman, K. B. *Biochemistry* (in the press).
- Morris, D. I., Speicher, L. A., Ruoho, A. E., Tew, K. D. & Seaman, K. B. *J. Biol. Chem.* **266**, 13377–13384 (1991).
- Rivoltini, L. *et al. Int. J. Cancer* **46**, 727–732 (1990).

- McCann, J. D., Li, M. & Welsh, M. J. *J. gen. Physiol.* **94**, 1015–1036 (1989).
- Giraldez, F., Valverde, M. A. & Sepúlveda, F. V. *Biochim. Biophys. Acta* **942**, 353–356 (1988).
- Okada, Y. & Hazama, A. *News physiol. Sci.* **4**, 238–242 (1989).
- Sheppard, D. N., Valverde, M. A., Giraldez, F. & Sepúlveda, F. V. *J. Physiol.* **433**, 663–676 (1991).
- Pastan, I. *et al. Proc. natn. Acad. Sci. U.S.A.* **85**, 4486–4490 (1988).
- Shen, D.-W. *et al. Science* **232**, 643–645 (1986).
- Tanaka, S. *et al. Biochem. biophys. Res. Commun.* **166**, 180–186 (1990).
- Suzuki, S., Tachibana, M. & Kaneko, A. *J. Physiol.* **421**, 645–662 (1990).
- Baas, F. *et al. Cancer Res.* **50**, 5392–5398 (1990).
- Fuerst, T. R., Niles, E. G., Studier, F. W. & Moss, B. *Proc. natn. Acad. Sci. U.S.A.* **83**, 8122–8126 (1986).
- Felgner, P. L. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 7413–7417 (1987).
- Karger, B. D. & Kornro, C. *Focus* **12**, 25–27 (1990).
- Laemmli, U. K. *Nature* **227**, 680–685 (1970).

ACKNOWLEDGEMENTS. We thank M. Gottesman for cell lines and antibodies and P. Borst for DNA; G. Smith and B. Moss for vaccinia stocks; J. Dempster for electrophysiological analysis software; and R. Boyd, M. Gibson, U. Gileadi, F. Giraldez, A. Harris, T. Keevit, J. O'Brien, P. Romano, C. Rucareanu, K. Seaman and S. Tucker for assistance and discussions. M.A.V. was supported by DGICYT (Spain) and M. D. by Gobierno Autonomo Canario and the British Council. This work was supported by the AFRC, the Cystic Fibrosis Research Trust and the Imperial Cancer Research Fund.

Exo1 and Exo2 proteins stimulate calcium-dependent exocytosis in permeabilized adrenal chromaffin cells

Alan Morgan & Robert D. Burgoyne

Department of Physiology, University of Liverpool, PO Box 147, Liverpool L69 3BX, UK

IN many cell types an increase in cytosolic calcium is the main signal for the exocytotic release of stored secretory components such as hormones and neurotransmitters. The site of action of calcium in exocytosis is not known, neither are the participating molecules^{1,2}. In the case of the intracellular membrane fusions that occur during transport through early stages of the secretory pathway, several cytosolic and peripheral membrane proteins are necessary^{3–6}. Permeabilized cells have been useful in understanding the requirements for calcium and nucleotides in regulated exocytosis^{7,8} and under certain conditions there is leakage of soluble protein components and run-down of the exocytotic response^{9–14}. This system can be used to identify the soluble proteins involved in exocytosis, one candidate in chromaffin cells being annexin II (calpactin)⁹. Here we use this assay to identify two other cytosolic protein factors that regulate exocytosis in permeabilized adrenal chromaffin cells, which we term Exo1 and Exo2. Exo1 from brain cytosol resolves on electrophoresis in SDS-polyacrylamide gels as a group of polypeptides of relative molecular mass ~30,000 and shares sequence homology with the 14–3–3 family of proteins. The ability of Exo1 to reactivate exocytosis is potentiated by protein kinase C activation and therefore Exo1 may influence the protein kinase C-mediated control of Ca²⁺-dependent exocytosis.

We used digitonin-permeabilized bovine adrenal chromaffin cells, which secrete catecholamines by exocytosis in response to micromolar calcium ion concentrations^{15,16}. Following permeabilization, these cells leak cytosolic proteins and secretory responsiveness runs down with a comparable half-time^{9,14}.

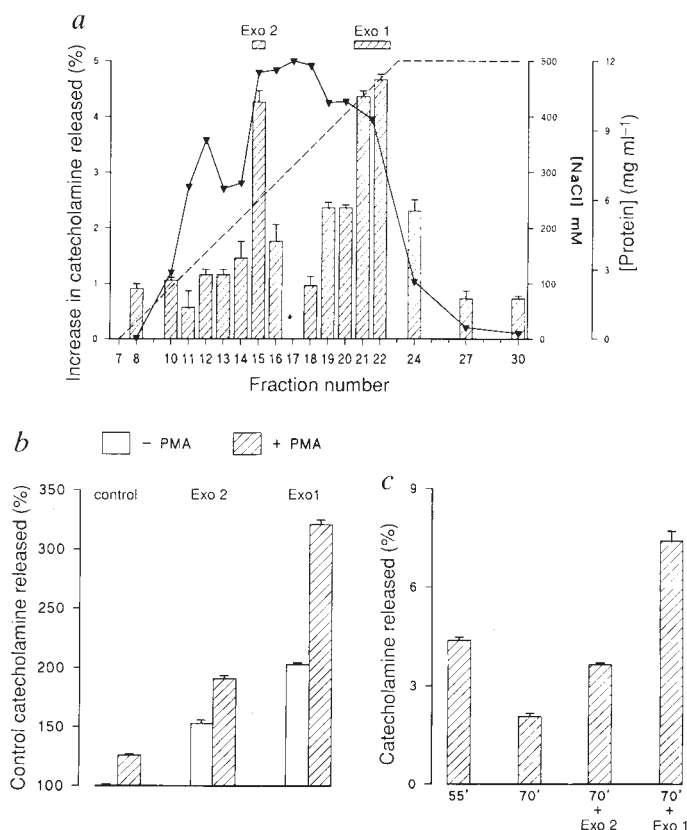
Secretory run-down can be retarded by incubation of the permeabilized cells with leaked proteins¹⁴, although the nature of the protein(s) responsible is unknown. In addition, we have found that annexin II, which has been implicated in exocytosis^{17,18}, will significantly retard run-down^{9,19} but cannot reactivate exocytosis after run-down²⁰. It seems that annexin II is required for exocytosis but that other essential components are lost from the cells. We have therefore used the permeabilized cell/run-down system to search for cytosolic proteins that are able to retard run-down and reactivate exocytosis.

Leaked protein preparations do not provide enough material for purification purposes and so whole-tissue cytosol fractions were used. Chromaffin cells were permeabilized with digitonin for 10 min, then incubated for a further 15 min in the absence of digitonin with or without cytosol fractions before challenge with 0 or 10 μ M Ca²⁺. Adrenal medullary cytosol retarded run-down, leading to a Ca²⁺-dependent increase in catecholamine secretion when cells were challenged 25 min after the start of permeabilization. This phenomenon was due to a trypsin- and N-ethylmaleimide-sensitive component. Bovine serum albumin at the same protein concentration could not replace cytosol. Depletion of annexin II from cytosol did not abolish its ability to retard run-down, as reported previously²¹. Annexin II does not bind to anion-exchange columns²², but we found that adrenal medullary cytosolic proteins capable of stimulating secretion bound to and were eluted from the anion exchanger Q-Sepharose. The same activities were detected in sheep brain cytosol (Fig. 1a), which was used for further characterization and purification. Among the proteins eluting from Q-Sepharose were two stimulatory factors (Exo1 and Exo2) and an inhibitory factor (denoted by our asterisk in Fig. 1a).

Activation of protein kinase C increases Ca²⁺-dependent exocytosis in chromaffin cells^{23–25}. Prior treatment of chromaffin cells with phorbol myristyl acetate (PMA) activates protein kinase C and prevents it from leaking out after permeabilization²⁶. PMA treatment gave only a small increase in secretion over and above that due to Exo2, but the stimulatory effect of the Exo1 fraction was markedly potentiated (Fig. 1b). It is not clear whether the residual secretion in control cells that was barely affected by PMA involves residual Exo1, Exo2 or some

FIG. 1 Fractionation on Q-Sepharose and characterization of factors in sheep brain cytosol that stimulate Ca^{2+} -dependent exocytosis in permeabilized chromaffin cells. **a**, Sheep brain cytosol was run on a Q-Sepharose column and bound proteins eluted with a salt gradient. The solid line shows protein concentration of fractions and the broken line the salt gradient. Fractions were tested for their ability to maintain responsiveness of cells to Ca^{2+} and the results are shown in the hatched bars as the increase in Ca^{2+} -dependent catecholamine release above controls as percentage of total cellular catecholamine. Two stimulatory factors were designated Exo1 and Exo2. Fraction 17 (asterisk) contained an inhibitory activity, which is not shown. Secretion from control cells was 6.4% of total catecholamine. **b**, Effect of treatment with PMA on the stimulatory activity of Exo1 and Exo2. Control cells were treated using the standard assay. Data are means $\pm$ s.e.m. ($n=4$) and are shown as percentage of secretion from control cells in response to $10 \mu\text{M}$ Ca^{2+} . Secretion from control cells was 8.2% of total catecholamine. **c**, Effect of Exo1 and Exo2 on Ca^{2+} -dependent secretion after long-term run-down. Cells were permeabilized for 55 min to allow extensive run-down and then incubated for 15 min without or with Exo1 (1.3 mg per well) or Exo2 (2.3 mg per well) before stimulation with $10 \mu\text{M}$ Ca^{2+} at 70 min. Data are shown as mean $\pm$ s.e.m. ($n=4$) and are expressed as a percentage of total cellular catecholamine.

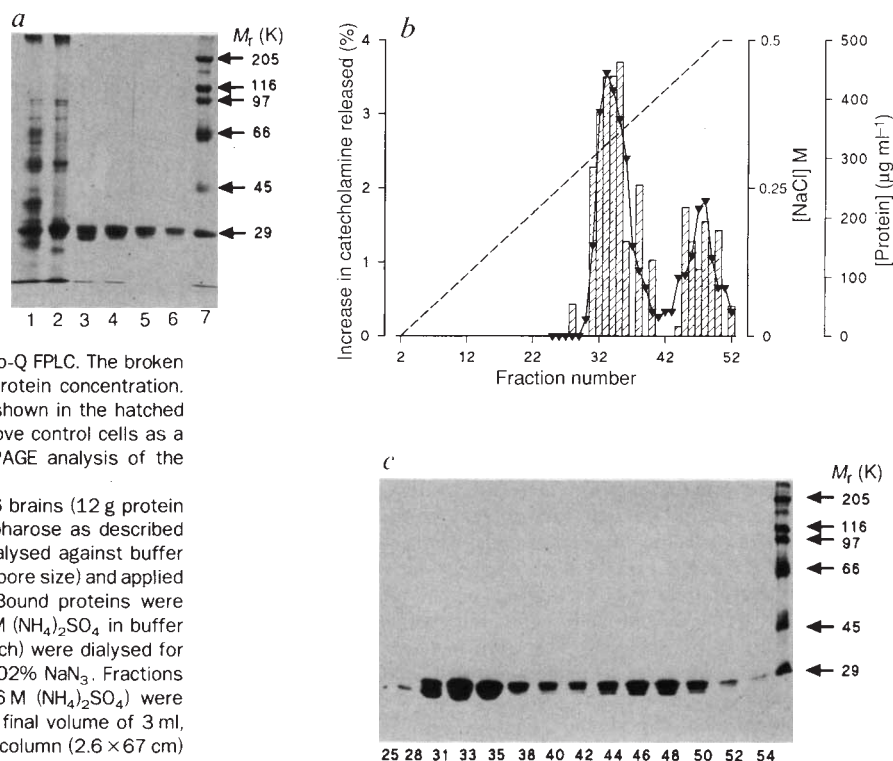
METHODS. Sheep brains were homogenized in 125 ml per brain of ice-cold buffer A (139 mM potassium glutamate, 2 mM ATP, 2 mM MgCl_2 , 5 mM EGTA, 20 mM PIPES, pH 6.5) containing 20 μM digitonin (Calbiochem), 1 mM DTT and 0.02% NaN_3 , centrifuged at 4,000g for 5 min, and the supernatant centrifuged at 100,000g for 60 min. The 100,000g supernatant was dialysed overnight against 10 volumes of buffer B (10 mM imidazole, 1 mM EGTA, 1 mM NaN_3 , 0.5 mM DTT, pH 7.3) applied to a 22-ml Q-Sepharose column (Pharmacia) and bound proteins eluted with a 48 ml 0–500 mM NaCl gradient in buffer B at 1 ml min^{-1} and 3-ml fractions collected. Eluted fractions were dialysed against buffer A with 1 mM DTT, 0.02% NaN_3 overnight. For standard testing of column fractions, chromaffin cells were isolated and cultured in 24-well trays⁹. Cultures were washed once with buffer A, permeabilized for 10 min in buffer A containing 20 μM digitonin (300 μl per well) and incubated for a further 15 min with column fractions or dialysis buffer (200 μl per well). Cultures were then challenged by replacement of buffer with buffer A either without calcium or with CaCl_2 added to a free Ca^{2+} concentration of $10 \mu\text{M}$ (300 μl per well). After 20 min, buffer was removed, centrifuged at 16,000g for 2 min and aliquots assayed fluorimetrically for released endogenous catecholamine. Total catecholamine left in the cells was deter-



mined by releasing it using 1% Triton X-100. All experiments were at room temperature. The protein concentration of fraction added was chosen to lie in the linear range of the assay. To investigate the effects of PMA treatment, 200 nM PMA was included in the permeabilization step and in the incubation with column fractions.

FIG. 2 Purification of Exo1. **a**, SDS-PAGE analysis of the fractions through the purification of Exo1. Lane 1, sheep brain cytosol; lane 2, Q-Sepharose Exo1 pool; lane 3, peak fractions from phenyl-Sepharose; lane 4, peak fractions from Sephacryl S-100 gel filtration; lane 5, first peak from Mono-Q (M1); lane 6, second peak from Mono-Q (M2); lane 7, M_r standards. The specific activities (per cent increase in secretion per mg protein) of the fractions in lanes 1–6 were 3.5, 14.3, 45.2, 88, 151 and 250, respectively. When the fractions were run on heavily overloaded gels, minor contaminating polypeptides were detected in fractions from phenyl-Sepharose and Sephacryl S-100, but these were removed on Mono-Q. **b**, Chromatography on Mono-Q FPLC. The broken line shows the salt gradient and the solid line the protein concentration. The effect of fractions on catecholamine release is shown in the hatched bars as the increase in Ca^{2+} -dependent secretion above control cells as a percentage of total cellular catecholamine. **c**, SDS-PAGE analysis of the fractions from Mono-Q (in b).

METHODS. Sheep brain cytosol was prepared from 16 brains (12 g protein in 21 buffer) and fractionated in 8 batches on Q-Sepharose as described in Fig. 1 legend. Active acidic Exo1 fractions were dialysed against buffer B containing 1M $(\text{NH}_4)_2\text{SO}_4$, millipore-filtered (0.22- μm pore size) and applied to a 22-ml phenyl-Sepharose column (Pharmacia). Bound proteins were eluted with a 96-ml decreasing salt gradient from 1M $(\text{NH}_4)_2\text{SO}_4$ in buffer B to buffer B at 1 ml min^{-1} . Eluted fractions (3 ml each) were dialysed for testing against buffer A containing 1 mM DTT and 0.02% NaN_3 . Fractions with peak activity (eluting between 0.25M and 0.06 M $(\text{NH}_4)_2\text{SO}_4$) were precipitated with 100% $(\text{NH}_4)_2\text{SO}_4$, resuspended to a final volume of 3 ml, run on a Sephacryl S-100HR (Pharmacia) gel filtration column (2.6 $\times$ 67 cm) equilibrated and eluted with buffer A containing 1 mM DTT and 0.02% NaN_3 at 2 ml min^{-1} , and 4-ml fractions collected. Fractions with peak activity (fractions 40 and 41) were dialysed against buffer B, applied to a 1-ml Mono-Q FPLC (Pharmacia) column, bound protein eluted with a 48-ml 0–500 mM NaCl gradient at 1 ml min^{-1} and 1-ml fractions collected. All



chromatography was at room temperature and fractions stored at -20°C after dialysis for testing. Fractions were dialysed and tested using PMA-treated cells as described in Fig. 1 legend.

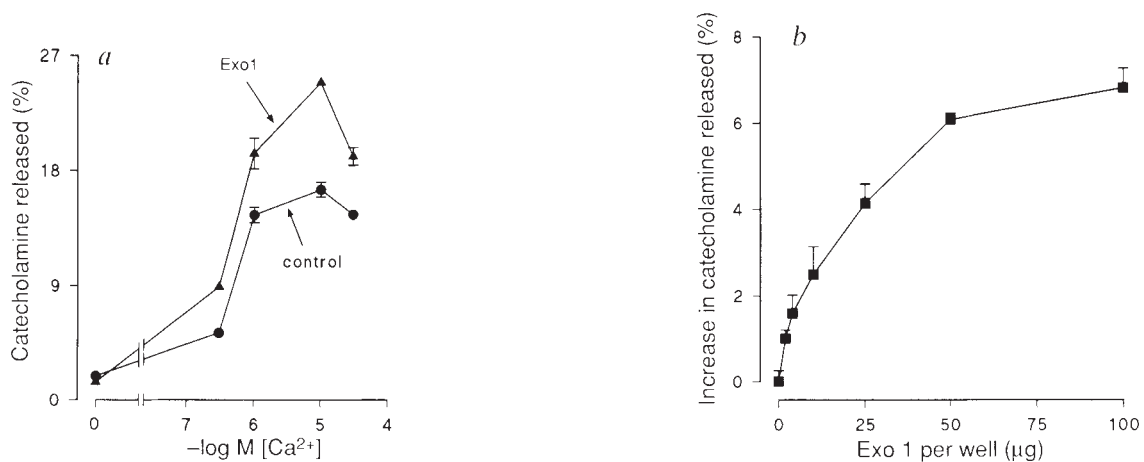


FIG. 3 Activity of purified Exo1. *a*, Ca²⁺-dependency of the stimulatory effect of Exo1 (113 µg per well) on secretion. *b*, Dose-dependence of the stimulatory effect of Exo1. Exo1 was purified through to the Mono-Q stage and

fractions from the two peaks pooled for testing. Cells were tested with PMA treatment using the standard protocol. Data are means $\pm$ s.e.m. ($n=4$).

other agent. All subsequent experiments used PMA-treated cells. We next tested the ability of Exo1 and Exo2 to reactivate exocytosis. Even after run-down of 55 min, incubation for 15 min with Exo1 reactivated exocytosis so that secretion was stimulated to levels above those before Exo1 addition (Fig. 1c). Exo2 retarded run-down but did not seem to reactivate secretion. Although Exo1 could reactivate exocytosis, this ability progressively decreased, indicating that Exo1 alone was not adequate for maintaining the response. The reduced effectiveness of Exo1 after long permeabilization times, coupled with its Ca²⁺-dependency and protein kinase C regulation, suggests that Exo1 acts on exocytosis rather than simply causing nonspecific granule lysis. To assess the ability of Exo1 to retard run-down, cells were permeabilized for 10 min and then incubated in the presence of Exo1 for a further 60 min. Secretion at this time in control cells was reduced to 31.6% of the initial control value; the response from Exo1-treated cells was maintained at 81% of the initial value, showing that run-down by Exo1 was considerably retarded. Exo1 and Exo2 together had additive effects on secretion.

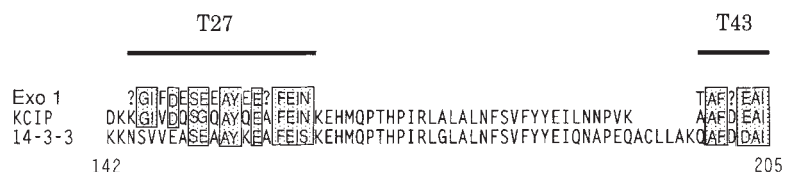
Gel filtration analysis of Exo2 gave a single peak corresponding to a relative molecular mass (M_r) of $\sim 44,000$ (44K). As Exo1 has a more marked effect on catecholamine secretion and may be important in exocytosis, we decided to purify this factor by hydrophobic interaction chromatography on phenyl-

Sephacrose, gel filtration and anion exchange on Mono-Q FPLC (Fig. 2a). This gave a 70-fold increase in specific activity through to the second Mono-Q peak. The activity ran as a single 70K peak on gel filtration, but ran as unusually broad peaks on phenyl-Sepharose and on Mono-Q. Many other polypeptides were resolved in single fractions on these columns. Exo1 activity correlated with a group of 3 closely migrating polypeptides corresponding to $M_r \sim 30K$ on SDS-PAGE. On the Mono-Q column, two peaks of activity contained these 30K polypeptides (Fig. 2b, c). Exo1 activity seems, therefore, to reside in one or more members of a family of closely related polypeptides. The 30K polypeptides were enriched early in the purification (Fig. 2a), indicating that they are relatively abundant in brain cytosol. The specific activity of M2 was higher than that of M1 after fractions had been pooled, although individual fractions had similar specific activities; this may be due to an increase in effectiveness of a mixture of Exo1 forms.

The stimulatory activity of purified Exo1 (combined fractions from the two Mono-Q peaks) is shown in Fig. 3. This activity is Ca²⁺-dependent (Fig. 3a) and detectable at microgram levels of protein. The activity of purified Exo1 is enhanced by prior treatment of cells with PMA and this is seen even at saturating levels of Exo1 (data not shown). Purified Exo1 gave half-maximal stimulation of secretion at 20 µg protein per well (Fig. 3b). This probably does not represent the true concentration at

FIG. 4 Sequence analysis (one-letter amino-acid code) of peptides derived from Exo1 compared with 14-3-3 and protein kinase C inhibitory proteins (KCIP). Tryptic peptides were generated from purified Exo1 (pooled Mono-Q fractions), separated by HPLC and sequenced. The sequences of two peptides (T27 and T43) are shown for comparison with the sequence for KCIP and the full sequence of 14-3-3. Homologous amino acids are boxed. The positions of residues in 14-3-3 are indicated at the bottom. Of 22 residues unambiguously identified from Exo1, 17 are identical to residues in 14-3-3 and KCIP, and the peptides would be preceded by lysines before trypsin digestion.

METHODS. Exo1 was digested with trypsin (substrate-to-enzyme ratio 30:1) at 37 °C overnight and the reaction stopped by addition of trifluoroacetic acid to 1%. The digest was separated by HPLC reverse-phase chromatography using a 0–60% acetonitrile gradient in 0.1% trifluoroacetic acid, and peptides were sequenced using an Applied Biosystems 471A gas-phase sequencer with on-line PTH-amino acid analyser.



exocytotic sites in the cells. Diffusion of proteins out of digitonin-permeabilized cells is relatively slow and diffusion of antibodies across chromaffin cells takes up to 16 minutes²⁷, suggesting that permeability to proteins is restricted. Therefore the effective concentration of Exo1 is likely to be much lower than that added to the cells. The increase in secretion due to purified Exo1 or impure fractions varied between different cell batches, which accounts for the differences in our results.

The features of Exo1 and Exo2 do not resemble those of other proteins proposed to participate in exocytosis or the early stages of the secretory pathway^{4-6,28}. The chromatographic properties of Exo1 are very similar to those of a family of proteins comprising the brain 14-3-3 protein²⁹⁻³¹ and a partially characterized group of 30K inhibitors of protein kinase C (refs 32, 33), which share a high degree of homology with 14-3-3. The 14-3-3 family of proteins, not previously implicated in exocytosis, comprises several widely distributed, homologous gene products^{29,30,33}, the functions of which are not clear. Peptide mapping indicated that the Exo1 polypeptides resolved on Mono-Q were similar across the gradient, so to assess whether Exo1 was related to the 14-3-3 family, we digested Exo1 with trypsin, separated the products by HPLC and sequenced two randomly selected peptides. The peptides showed a high degree of homology with 14-3-3 and the protein kinase C inhibitory proteins (Fig. 4), indicating that Exo1 is a member of the 14-3-3 family. The conserved sequences of the 14-3-3 proteins seem to be unique, apart from a domain that is very like the conserved C terminus of the annexins, in particular annexin II (refs 32, 33). As annexin II stimulates exocytosis in chromaffin cells^{9,19,20}, it is intriguing that two proteins that stimulate Ca^{2+} -dependent exocytosis in chromaffin cells should share this conserved domain.

In common with the process of membrane fusion in the early stages of the secretory pathway^{4-6,28}, exocytosis may require several proteins, including protein kinase C, annexin II, Exo1, Exo2, as well as other soluble and membrane²⁷ proteins. The ability of Exo1 to reactivate exocytosis following run-down in permeabilized cells, together with its potentiation by protein kinase C activation, suggest that this protein is a key factor in Ca^{2+} -dependent exocytosis. □

Defective anion transport activity of the abnormal band 3 in hereditary ovalocytic red blood cells

Ann E. Schofield, David M. Reardon*
& Michael J. A. Tanner†

Department of Biochemistry, School of Medical Sciences,
University of Bristol, Bristol BS8 1TD, UK

* Department of Haematology, Addenbrookes Hospital,
Cambridge CB2 2QQ, UK

† To whom correspondence should be addressed

HEREDITARY ovalocytosis is common in some areas of Melanesia and South East Asia where malaria is endemic. These red cells resist invasion by malarial parasites *in vitro*^{1,2} and ovalocytic individuals are less parasitized than normal³. This has been attributed to the greater rigidity of ovalocytic red cells^{4,5}. It has been suggested that South East Asian ovalocytosis results from the heterozygous presence of an altered membrane anion transporter (band 3)^{6,7}. We have used the polymerase chain reaction to clone the abnormal band 3 complementary DNA from an ovalocytic of Indian origin⁸ and found two changes from the normal protein: a point mutation (Lys 56 → Glu) and the deletion of the sequence AFSPQVLAA (residues 400–408), but no evidence for an N-terminal extension⁷. The deletion is also found in the abnormal band 3 of South East Asian ovalocytes⁹ and seems to be responsible for the unusual properties of the ovalocytic red cell. We show here that the membrane domain of the abnormal ovalocyte band 3 has a substantially altered structure and that the protein is defective in anion transport activity. The changed transport properties of the red cells may have a role in the reduced parasitaemia of ovalocytic individuals.

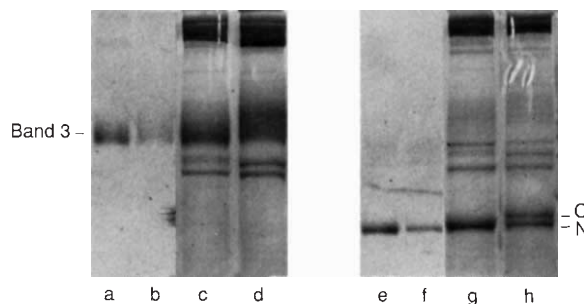


FIG. 1 Covalent binding of [³H]H₂DIDS to band 3 of normal and ovalocytic red cells. Lanes a–f, autoradiographs of SDS–polyacrylamide gels showing membranes from [³H]H₂DIDS-labelled normal (lane a) and ovalocyte (b) red cells, and from [³H]H₂DIDS-labelled chymotrypsin-treated normal (e) and ovalocyte (f) red cells. Lanes c–h, protein-stained gels of identical samples run in parallel of the membranes from [³H]H₂DIDS-labelled normal (lane c) and ovalocyte (d) red cells; [³H]H₂DIDS-labelled chymotrypsin-treated normal (g) and ovalocyte (h) red cells. The N-terminal 60K membrane-bound chymotryptic band 3 fragment from normal (N) and ovalocyte (O) are arrowed. The ovalocyte sample was a Mauritian of Indian origin who is described in ref. 8.

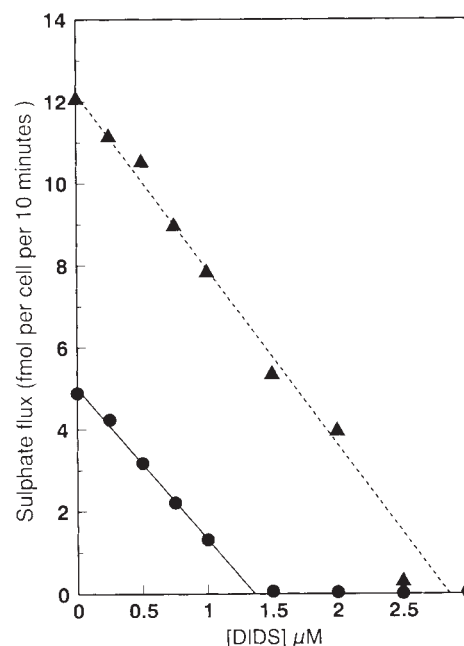
METHODS. Red cells washed with 0.1 M sodium phosphate, pH 7, were incubated at 10% haematocrit for 60 min with 5 μM [³H]H₂DIDS, and then washed with 0.5% bovine serum albumin in 0.1 M sodium phosphate, pH 7. The band 3 in a portion of the labelled cells was cleaved by chymotrypsin treatment (1 mg ml⁻¹ for 60 min at 37 °C). Membranes were prepared and separated by SDS–PAGE¹⁷ on gels containing 8% acrylamide. Identical volumes of samples were run on two parallel gels. One gel was stained with Coomassie blue and the other was treated for fluorography.

Received 5 August; accepted 5 December 1991.

1. Almers, W. A. *Rev. Physiol.* **52**, 607–624 (1990).
2. Burgoyne, R. D. *Biochim. biophys. Acta* **1071**, 174–202 (1991).
3. Balch, W. E. *J. Biol. Chem.* **264**, 16965–16968 (1989).
4. Clary, D. O., Griff, I. C. & Rothman, J. E. *Cell* **61**, 709–721 (1990).
5. Wattenberg, B. W., Hiebsch, R. R., LeCureux, L. W. & White, M. P. *J. Cell Biol.* **110**, 947–954 (1990).
6. Wilson, D. W. *et al. Nature* **339**, 355–359 (1989).
7. Gomperts, B. D. *A. Rev. Physiol.* **52**, 591–606 (1990).
8. Knight, D. E., von Grafenstein, H. & Athayde, C. M. *Trends Neurosci.* **12**, 451–458 (1989).
9. Ali, S. M., Geisow, M. J. & Burgoyne, R. D. *Nature* **340**, 313–315 (1989).
10. Holz, R. W., Bittner, M. A., Peppers, S. C., Senter, R. A. & Eberhard, D. A. *J. Biol. Chem.* **264**, 5412–5419 (1989).
11. Howell, T. W., Kramer, I. M. & Gomperts, B. D. *Cell Signal.* **1**, 157–163 (1989).
12. Koffer, A. & Gomperts, B. D. *J. Cell Sci.* **94**, 585–591 (1989).
13. Martin, T. F. J. & Waient, J. H. *J. Biol. Chem.* **264**, 10299–10308 (1989).
14. Sarafian, T., Aunis, D. & Bader, M.-F. *J. Biol. Chem.* **262**, 16671–16676 (1987).
15. Dunn, L. A. & Holz, R. W. *J. Biol. Chem.* **258**, 4989–4993 (1983).
16. Schafer, T., Karli, U. O., Gratwohl, E. K.-M., Schweizer, F. E. & Burger, M. M. *J. Neurochem.* **49**, 1696–1707 (1987).
17. Drust, D. S. & Creutz, C. E. *Nature* **331**, 88–91 (1988).
18. Nakata, T., Sobue, K. & Hirokawa, N. *J. Cell Biol.* **110**, 13–25 (1990).
19. Ali, S. M., Geisow, M. J. & Burgoyne, R. D. *Cell Signal.* **2**, 265–276 (1990).
20. Burgoyne, R. D. & Morgan, A. *Biochem. Soc. Trans.* **18**, 1111–1114 (1990).
21. Wu, Y. N. & Wagner, P. D. *FEBS Lett.* **282**, 197–199 (1991).
22. Glenney, J. R., Tack, B. & Powell, M. A. *J. Cell Biol.* **104**, 503–511 (1987).
23. Burgoyne, R. D., Morgan, A. & O'Sullivan, A. *J. FEBS Lett.* **238**, 151–155 (1988).
24. Knight, D. E. & Baker, P. F. *FEBS Lett.* **160**, 98–100 (1983).
25. Pocotte, S. L. *et al. Proc. natn. Acad. Sci. U.S.A.* **82**, 930–934 (1985).
26. Terbush, D. R. & Holz, R. W. *J. Biol. Chem.* **261**, 17099–17106 (1986).
27. Schweizer, F. E. *et al. Nature* **339**, 709–712 (1989).
28. Block, M. R., Glick, B. S., Wilcox, C. A., Wieland, F. T. & Rothman, J. E. *Proc. natn. Acad. Sci. U.S.A.* **85**, 7852–7856 (1988).
29. Boston, P. F., Jackson, P. & Thompson, R. J. *J. Neurochem.* **38**, 1475–1482 (1982).
30. Ichimura, T. *et al. Proc. natn. Acad. Sci. U.S.A.* **85**, 7084–7088 (1988).
31. Yamanuchi, T., Nakata, H. & Fujisawa, H. *J. Biol. Chem.* **256**, 5404–5409 (1981).
32. Toker, A., Ellis, C. A., Sellers, L. A. & Aitken, A. *Eur. J. Biochem.* **191**, 421–429 (1990).
33. Aitken, A., Ellis, C. A., Harris, A., Sellers, L. A. & Toker, A. *Nature* **344**, 594 (1990).

ACKNOWLEDGEMENTS. This work was supported by a project grant from the Medical Research Council and the Wellcome Trust to R.D.B. We thank J. Smith and M. Wilkinson for help with protein sequencing.

FIG. 2 DIDS titration of sulphate influx into normal ($\blacktriangle$) and ovalocytic ($\bullet$) red cells. Red cells were washed three times in 84 mM sodium citrate, 1 mM EGTA, pH 6.5. The influx of [35 S]sulphate into the red cells at 10% haematocrit was measured in a medium containing 4 mM sodium sulphate, 84 mM sodium citrate, 1 mM EGTA pH 6.5. Influx was measured after 10 min at 30 °C in the presence of different concentrations of DIDS. Studies of the time course of sulphate uptake showed that uptake was still linear at this time. The number of cells in each sample was counted in a haemocytometer and the data are presented as sulphate influx per red cell. We confirmed that, despite their different shapes, each ovalocyte cell contained the same number of band 3 molecules as a normal red cell, by quantitation of the protein staining of band 3 on SDS-PAGE of membranes derived from equal numbers of ovalocytes and normal cells. The ovalocyte sample was as described for Fig. 1 and the normal blood was taken at the same time. This experiment gives an estimate of both the relative transport activity (from the intercept with the sulphate flux) and relative number of DIDS-binding sites (whether reversible or irreversible) in the two cell types from the extrapolated intercept with the DIDS concentration. DIDS has a very high affinity for normal band 3, both in its fast initial binding as a reversible inhibitor ($K_i = 0.03 \mu\text{M}$; ref. 18) and in its slower subsequent action as an irreversible inhibitor ($K_i = 0.04 \mu\text{M}$; ref. 19). The concentration of DIDS-binding band 3 protein sites in this experiment is $\sim 2 \mu\text{M}$ for the normal cells and $1 \mu\text{M}$ for the ovalocytes. Under these conditions, in which the concentration of protein binding sites is very much higher than the K_i of the inhibitor²⁰, DIDS stoichiometrically binds and inhibits the band 3 in the cells until inhibition is almost complete, so that DIDS titrates band 3 until this point²¹. The absolute values for the number of DIDS inhibitory sites are liable to error because of impurities in the DIDS, and because of differences in the packed cell volume of the ovalocytes and normocytes resulting from their different shapes, but the results show that within the range of inhibitor concentrations used, the number of DIDS binding sites in ovalocytes is reduced to about half that in normal cells. The displacement of the two parallel lines in the figure is a



result of different numbers of active protein transport sites and DIDS binding sites being present in equal numbers of normal and ovalocytic red cells, and is not because of a difference in the affinity of DIDS binding to the band 3 of the two cell types.

The deletion in ovalocytic band 3 spans the junction of the N-terminal cytoplasmic domain and four amino acids of the first membrane span¹⁰. The latter probably acts as an uncleaved signal sequence¹¹ and is important for the assembly of the membrane domain. We used di-isothiocyanato-dihydrostilbene disulphonate (H_2DIDS), a high-affinity anion-transport inhibitor which covalently reacts with band 3, as a probe for structural changes in ovalocytic band 3. Ovalocytes and normal red cells were treated with [^3H] H_2DIDS and the membrane proteins separated by SDS-polyacrylamide gel electrophoresis (Fig. 1, lanes a-d). Much less radioactivity was found in the band 3 of the ovalocytes than normal cells. The radioactivity in the ovalocytes was associated with the normal band 3 (which migrates in the front portion of the band 3 region), rather than with the more slowly migrating abnormal band 3 of these heterozygous cells. The 60K N-terminal fragments (relative molecular mass, 60,000) generated by extracellular chymotrypsin cleavage of normal and ovalocytic band 3 are clearly separated by SDS-PAGE⁷ and retain the H_2DIDS binding site. Figure 1 (lanes e-h) shows that [^3H] H_2DIDS was found only in the 60K chymotryptic fragment derived from the normal band 3 in the ovalocytes, confirming the lack of reactivity of the abnormal band 3.

As the deletion is at the cytoplasmic boundary of the first membrane span and H_2DIDS binds in an extracellular portion of the fifth membrane span^{10,11}, these results were suggestive of a major structural change in the membrane domain of the abnormal protein. We therefore examined the band 3-mediated anion transport activity of the ovalocytic red cells (Fig. 2; three experiments gave similar results) and found that the ovalocytes had 41%, 34% and 46% of the sulphate transport activity of control red cells. Figure 2 also shows that the abnormal protein does not bind DIDS reversibly or irreversibly at the concentrations used. The abnormal band 3 in the ovalocytes seems to have little or no sulphate transport activity. The altered structure of the membrane domain may change the association properties

of ovalocytic band 3 and contribute to the unusual membrane mechanical properties of the cells.

The decreased anion transport activity of ovalocytic red cells may also affect intraerythrocytic malarial parasite growth in ovalocytic red cells. Cabantchik *et al.*¹² covalently modified normal red cells with DIDS and showed a reduction of intraerythrocyte *P. falciparum* growth *in vitro* in parallel with the reduction in red cell anion-transport activity. They suggested that inhibition of phosphate or sulphate entry through band 3 may limit intraerythrocytic growth of the rapidly developing parasites¹².

Outside the red cell the erythroid band 3 gene product is found only in the human kidney, where it is involved in bicarbonate reabsorption and acid secretion¹³. Defects in this kidney function, as well as red cell function, may make homozygosity for the variant band 3 lethal. There may be physiological consequences other than impaired acid-base homeostasis even for heterozygotes, as band 3-catalysed bicarbonate flux limits bicarbonate transport in the microcapillary system *in vivo*¹⁴. Ovalocytic individuals might be expected to have reduced bicarbonate transport and impaired oxygen uptake and release by the blood, which would be particularly significant during exercise and at altitude. It is interesting that ovalocytosis in Papua New Guinea occurs in the coastal lowlands but is rare in the adjacent highlands, a distribution attributed solely to the incidence of malaria in the coastal lowlands^{15,16}.

Note added in proof: We have recently expressed the ovalocytic band 3 cRNA in *Xenopus* oocytes and confirmed the protein lacks anion transport activity (J. D. Groves, A.E.S. and M.J.A.T., unpublished experiments). $\square$

Received 5 August; accepted 22 November 1991.

1. Kidson, C., Lamont, G., Saul, A. & Nurse, G. T. *Proc. natn. Acad. Sci. U.S.A.* **78**, 5829-5832 (1981).
2. Hadley, T. *et al. J. clin. Invest.* **71**, 780-782 (1983).
3. Cattani, J. A., Gibson, F. D., Alpers, M. P. & Crane, G. G. *Trans. R. Soc. Trop. Med. Hyg.* **81**, 705-709 (1987).

4. Saul, A., Lamont, G., Sawyer, W. H. & Kidson, C. *J. Cell. Biol.* **98**, 1348–1354 (1984).
5. Mohandas, N., Lie-Injo, L. E., Friedman, M. & Mak, J. W. *Blood* **63**, 1385–1392 (1984).
6. Liu, S. C. *et al.* *New Engl. J. Med.* **323**, 1530–1538 (1990).
7. Jones, G. L., Edmunds, H. M., Wesche, D. & Saul, A. *Biochim. biophys. Acta* **1096**, 33–40 (1991).
8. Schofield, A. E. *et al.* *J. molec. Biol.* (in the press).
9. Tanner, M. J. A., Bruce, L., Martin, P. G., Rearden, D. M. & Jones, G. L. *Blood* **78**, 2785–2786 (1991).
10. Tanner, M. J. A., Martin, P. G. & High, S. *Biochem. J.* **256**, 703–712 (1988).
11. Kopito, R. R. & Lodish, H. F. *Nature* **316**, 234–238 (1985).
12. Cabantchik, Z. I., Kutner, S., Krugliak, M. & Ginsburg, H. *Molec. Pharmac.* **23**, 92–99 (1983).
13. Wagner, S., Vogel, R., Lietzke, R., Koob, R. & Drenckhahn, D. *Am. J. Physiol.* **253**, F213–F221 (1987).
14. Weith, J. O., Anderson, O. S., Brahm, J., Bjerrum, P. J. & Borders, C. L. *Phil. Trans. R. Soc. B* **299**, 383–399 (1982).
15. Amato, D. & Booth, P. B. *Papua New Guinea Med. J.* **20**, 26–32 (1977).
16. Serjeantson, S., Bryson, K., Amato, D. & Babona, D. *Hum. Genet.* **37**, 161–167 (1977).
17. Laemmli, U. K. *Nature* **227**, 680–681 (1970).
18. Janas, T., Bjerrum, P. J., Brahm, J. & Weith, J. O. *Am. J. Physiol.* **257**, C601–C606 (1989).
19. Pimplikar, S. W. & Reithmeier, R. A. F. *J. biol. Chem.* **261**, 9770–9778 (1986).
20. Shearman, M. S. & Halestrap, A. P. *Biochem. J.* **223**, 673–676 (1984).
21. Halestrap, A. P. *Biochem. J.* **156**, 193–207 (1976).

ACKNOWLEDGEMENTS. We thank R. Poole and A. Halestrap for their help with anion transport assays. A.E.S. received a Wellcome Prize Studentship.

Mutations in the *Caenorhabditis elegans unc-4* gene alter the synaptic input to ventral cord motor neurons

J. G. White, E. Southgate & J. N. Thomson

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

IDENTIFICATION of the genes orchestrating neurogenesis would greatly enhance our understanding of this process. Genes have been identified that specify neuron type (for example *cut*¹ and *numb*² in *Drosophila* and *mec-3* in *Caenorhabditis elegans*³) and process guidance (for example, *unc-5*, *unc-6* and *unc-40* in *C. elegans*⁴ and the *fas-1* gene of *Drosophila*⁵). We sought genes defining synaptic specificity by identifying mutations that alter synaptic connectivity in the motor circuitry in the nematode *C. elegans*. We used electron microscopy of serial sections⁶ to reconstruct the ventral nerve-cords of uncoordinated (*unc*) mutants^{7,8} that have distinctive locomotory choreographies. Here we describe the phenotype of mutations in the *unc-4* gene in which a locomotory defect is correlated with specific changes in synaptic input to a subset of the excitatory VA motor neurons, normally used in reverse locomotion. The circuitry alterations do not arise because of the inaccessibility of the appropriate synaptic partners, but are a consequence of changes in synaptic specificity. The VA motor neurons with altered synaptic inputs are all lineal sisters of VB motor neurons; the VA motor neurons without VB sisters have essentially the same synaptic inputs as in wild-type animals. The normal function of the wild-type allele of *unc-4* may thus be to invoke the appropriate synaptic specificities to VA motor neurons produced in particular developmental contexts.

When adult *unc-4* mutants try to move backwards, either as part of their normal foraging behaviour or when provoked by a tap on the head, the mid-body bends into a tight curve with the dorsal side innermost (Fig. 1). But the head and tail are still capable of dorsal as well as ventral flexures. Consequently, *unc-4* mutants are often seen in an unusual 'Ω' configuration with the mid-body bent dorsally and the head and tail bent ventrally. First-stage (L1) larvae of *unc-4* mutants are normal, the characteristic uncoordinated phenotype only developing after the L1–L2 moult, corresponding to the time when several additional classes of motor neuron are added to those present in the newly hatched L1 larva⁸.

The ventral cord motor circuitry was reconstructed in three *unc-4(e120)* individuals, two in the anterior cord regions and one in the posterior cord. The general organization of the neuropil was indistinguishable from wild-type animals. All neural processes followed their characteristic trajectories through the neuropil and consequently were adjacent (and hence synaptically accessible) to the same set of neighbouring processes as their wild-type counterparts. All the classes of motor neuron in the ventral cord were identified (that is the VA, VB, DA, DB, VD, DD, AS and VC neurons⁶) and their cell bodies were found in their characteristic locations. With the exception of some VA motor neurons, all ventral cord motor neurons had the same synaptic inputs and synapsed onto the same targets as their wild-type counterparts.

There are 12 VA motor neurons in the wild-type ventral cord (VA1–VA12) innervating ventral muscles. The nine DA motor neurons have the same synaptic input as VA motor neurons and innervate dorsal muscles. The A-type motor neurons (VA and DA) have characteristic anteriorly directed axons⁶ and are used for reverse locomotion⁷. Forward locomotion depends on the B-type motor neurons (VB and DB) having similar morphologies to A-type neurons, except that they are mirror images with posteriorly directed axons⁶. Both A-type and B-type neurons are probably excitatory⁸ but receive synaptic input from different sets of interneurons⁶.

In reconstructions of *unc-4(e120)* animals, VA1, VA11 and VA12 neurons were similar to their wild-type counterparts, as were all the reconstructed DA motor neurons. By contrast, each of the VA neurons that was reconstructed in the middle region of the ventral cord (VA2, VA3 and VA10) was miswired (Figs 2, 3 and 4). Input from interneurons that usually synapse onto VA neurons (gap junctions from AVA interneurons, chemical synapses from AVA, AVD and AVE interneurons⁶) was replaced with input characteristic of B-type neurons (gap junctions from AVB interneurons⁶). The VA neurons that have altered synaptic input in *unc-4(e120)* are nonetheless morphologically indistinguishable from the corresponding VA motor neurons in the wild type and have the same post-synaptic specificity as wild-type VA neurons (they innervate ventral body muscles (Fig. 2)). The neuroanatomical phenotype of *unc-4(e120)* mutants was therefore quite well defined: a subset of the VA motor neurons had

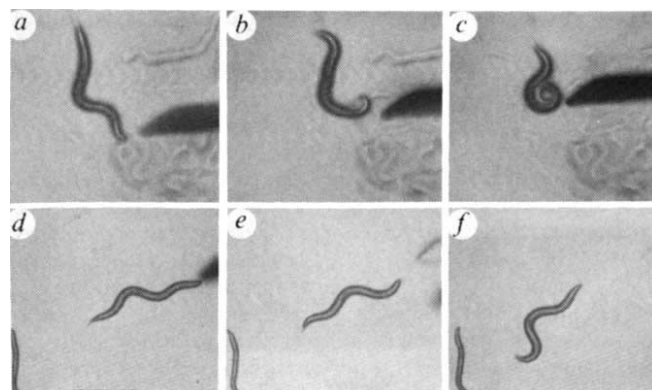


FIG. 1 Sequence of frames taken from video recordings comparing the behaviour of *unc-4(e120)* mutants with wild-type animals after being provoked to move backwards by tap on the head. *unc-4(e120)* animals respond to this stimulus by abruptly curling up their bodies, dorsal side innermost yet leaving their tails free (a–c). Often head and tail movements can be seen in these situations while the body is locked into a tight immobile bend. Wild-type animals will promptly move backwards when touched on the head by propagating forward directed waves along their bodies (d–f). Interval between frames, 1s. Scale bar, 500 μ m.

altered presynaptic specificity, that is they did not have their normal synaptic inputs, but rather received synaptic input normally characteristic of B-type neurons.

The behavioural phenotype of *unc-4(e120)* is consistent with the anatomical phenotype and with what is known about the function of the various classes of motor neuron in the ventral cord of *C. elegans*. B-type motor neurons receive their synaptic input from a common set of interneurons and are necessary for normal forward locomotion⁹. Similarly, all A-type motor neurons receive their synaptic input from a different set of interneurons and are necessary for reverse locomotion. In *unc-4(e120)* animals most or all of the VA neurons in the mid-body region probably receive their synaptic input from the interneurons associated with forward rather than reverse locomotion (Fig. 4). Therefore when *unc-4(e120)* animals try to move backwards by activating the appropriate interneurons, the VA neurons with altered synaptic inputs will not be activated. Ventral innervation of the mid-body will be consequently lacking, leading to a dorsal-ventral imbalance with only the dorsal side being innervated (by the DA neurons). This causes the unilateral activation of dorsal muscles bending the body with the ventral side outermost. We attribute the dorsal-ventral flexures of the head and tail that are seen to the fact that the VA neurons at either end of the cord (VA1, 11 and 12) receive their normal synaptic inputs. Although the VA4-9 neurons in the mid-body

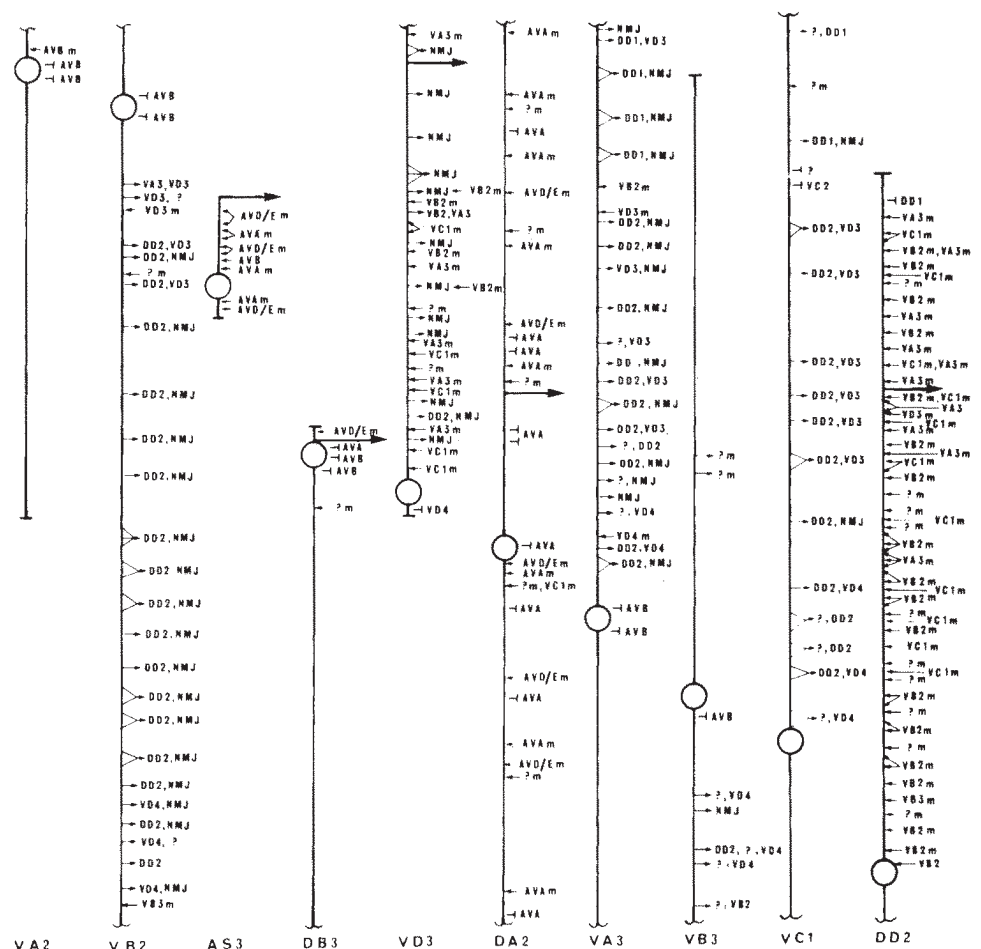
region have not been reconstructed, most or all are probably also transformed in the manner of VA2, 3 and 10 as the bend in the mid-body region never has a kink that would indicate that some of the VA motor neurons in the central region were functioning normally.

Forward locomotion in *unc-4(e120)* animals is slightly uncoordinated. We have not detected any changes in the B-type motor neurons that were reconstructed and therefore attribute the mild retardation of forward movement to interference from the inappropriately activated VA2-10 neurons.

The VA and VB motor neurons are produced postembryonically and make their synaptic connections during the L1-L2 moult (our unpublished observations). It is therefore likely that the uncoordinated phenotype appears when the misconnected VA neurons become functional.

An unusual aspect of the *unc-4(e120)* anatomical phenotype is that only a subset of VA motor neurons is affected. A possible explanation for this is that the alleles studied are leaky and that a null allele might cause all the A-type neurons to be affected. This seems unlikely, as the *e120* allele is identical behaviourally to the *wd1* allele in which a portion of the *unc-4* coding sequence is deleted¹⁰. Both alleles have the characteristic *unc-4* behavioural phenotype with the ends of the body able to make dorsal-ventral flexures while the mid-body is tightly curled up. This indicates that the VA1, 11 and 12 and all the DA motor

FIG. 2 Reconstruction of a region of the nervous system in the anterior ventral cord of an *unc-4(e120)* individual. Electron microscopy and serial section reconstructions were done as described⁶. Vertical lines, processes of axons (anterior is up); horizontal lines with large arrows, commissures leaving ventral cord; large circles, cell bodies; small arrows leaving a process, chemical synapses made to other neurons; small arrows coming into a process, synapses from other neurons (m indicates that the process was one of two or more post-synaptic elements at the synapse); 'T' endings, gap junctions; NMJ, neuromuscular junctions; ?, unidentified processes. Neurons VB2, AS3, DB3, VD3, DA2, VB3, VC1 and DD2 are essentially the same as those in wild-type animals (see ref. 6). But neurons VA2 and VA3 completely lack the characteristic synaptic inputs from AVA, AVD and AVE. These inputs can be seen on the normal-looking DA2 neuron which is unaffected by mutations in *unc-4* gene. But the VA2 and VA3 neurons make gap junctions to AVB interneurons in the vicinity of their cell bodies. These are the characteristic connections made onto B-type neurons by interneurons and can be seen on VB2, DB2 and VB3 which are very similar to their wild-type counterparts (the gap junction made by AVA onto DB3 is not usually seen but DB neurons in other regions do not make these connections). The predominant synaptic input from interneurons to type-B motor neurons is derived from AVB through gap junctions; however, there are also chemical synapses made by PVC interneurons to a random subset of B-type neurons in the wild type⁶. Although these connections were not seen on the *unc-4(e120)* VA neurons shown, two synapses from PVC were seen on a mutant VA10. Thus it appears from the reconstructions that VA2, VA3 and VA10 have switched both their gap junction and chemical presynaptic specificities to those appropriate to B-class neurons.



Behavioural observations suggest that VA4-9 are also transformed in a similar manner (see text). The synaptic output of VA and VB neurons in both wild-type⁶ and *unc-4(e120)* mutants is predominantly directed to ventral body muscles and D-type motor neurons. The *unc-4(e120)* mutation does not therefore seem to affect the postsynaptic specificity of any of the A-type neurons.

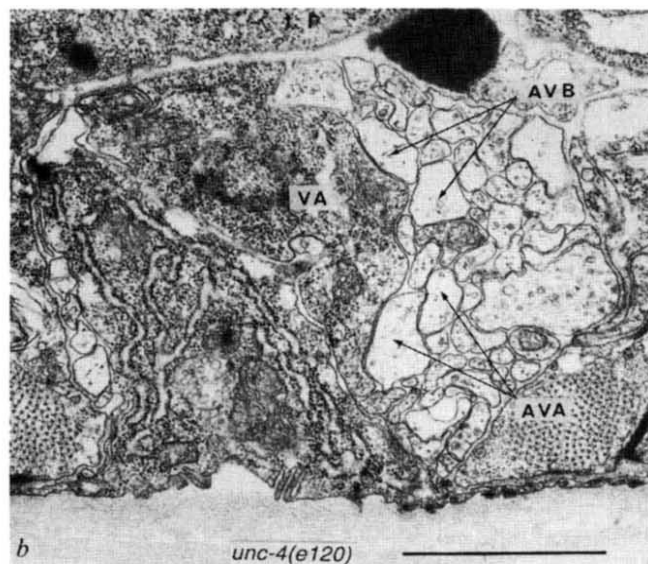
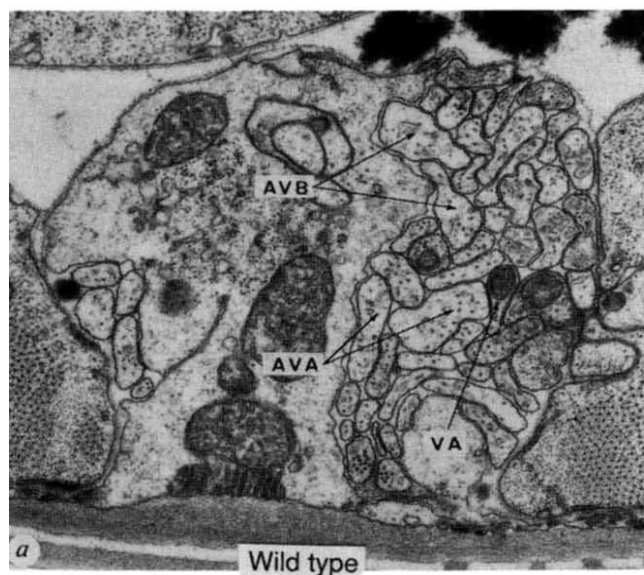


FIG. 3 Electron micrographs showing gap junction connections made by VA neurons onto interneurons in a wild-type animal (a) and an *unc-4(e120)* mutant (b). In the wild-type animals gap junctions are seen between VA motor neurons and AVA interneurons along the length of the VA process away from the cell body (a). In the mutant, VA neurons in the mid-body make gap junctions to AVB interneurons instead of AVA. These connections

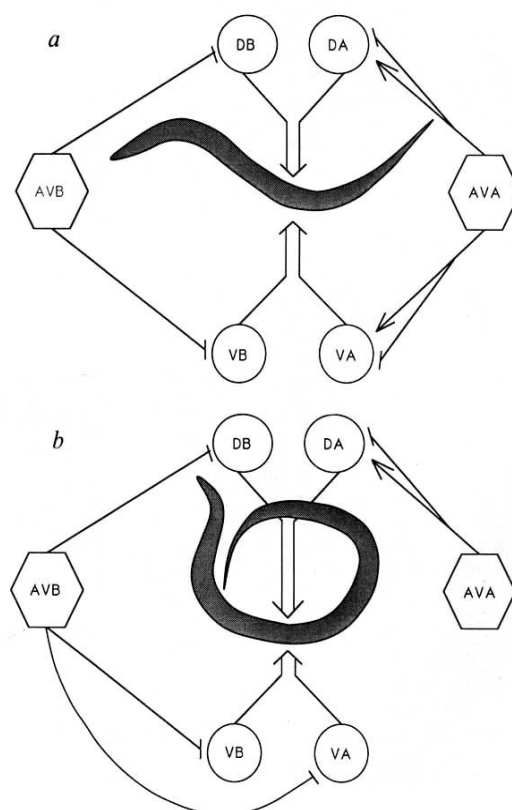
are seen in the vicinity of the cell body, as is the case for gap junctions between AVB interneurons and VB motor neurons in both wild-type and *unc-4(e120)* mutants⁶. Both AVA and AVB interneuron classes are equally accessible to the cell body regions of A-type neurons (b) indicating that the altered connections are a consequence of a change in synaptic specificity. Scale bar, 1 μ m.

neurons function normally in these animals and therefore do not require *unc-4* gene activity.

Why is only a subset of the A-type neurons affected in *unc-4* mutants? It may be significant that the set of modified VA neurons seen in *unc-4* mutants all share a similar lineal history. All VA and VB neurons are produced postembryonically along

with several other neuron types from a sequence of neuroblasts¹¹. The transformed VA neurons in *unc-4* mutants are all sisters to VB neurons. The unaffected VA11 and VA12 neurons are sisters to apoptosis and the unaffected VA1 neuron is sister to a VD motor neuron. The DA motor neurons are all produced embryonically but none are sisters of B-type neurons¹². Thus only the

FIG. 4 Summary of the connectivity of the excitatory motor neurons in the ventral cord of wild-type animals (a) and *unc-4(e120)* mutants (b). Arrows, chemical synapses; 'T' endings, gap junctions; circles, motor neuron classes; hexagons, interneuron classes. Forward locomotion is mediated by the B-type motor neurons driven by a specific set of interneurons of which AVB is the major class. Reverse motion is mediated by the A-type neurons driven by a different set of interneurons of which AVA is the major class⁶. The VA motor neurons in the mid-body of *unc-4(e120)* mutants have synaptic input that is characteristic of B-type neurons. When *unc-4* mutants try to move backward the body coils up with the dorsal side innermost (Fig. 1). This is consistent with the neuroanatomical changes that were seen in these mutants as the ventral body regions receive no synaptic input when animals try to reverse, the resulting imbalance between the ventral and dorsal sides causing the animals to coil up in this situation.



A-type neurons that are sisters of B-type neurons are transformed in *unc-4* mutants, and the transformation is such that they receive the same synaptic inputs as their sisters.

One possible interpretation of the *unc-4* loss-of-function phenotype is that those VA neurons that are produced as sisters to VB neurons become transformed into VB neurons. Similar transformations of the *C. elegans* neurons HSN to their lineal sisters PHB has also been seen in *ham-1* mutants¹³. The *unc-4* may then be thought of as a cell-identity gene. But the transformation is not complete as the polarity of axon growth of the transformed VA neurons is not appropriate for a VB. Axon polarity is likely to be an important, class-specific feature of A- and B-type motor neurons. When the commissures from DA

and DB neurons reach the dorsal mid-line during embryogenesis, those from DA all turn to run anteriorly, whereas those from DB all turn to run posteriorly in the absence of any obvious influencing factors apart from their interactions with the dorsal hypodermis¹⁴.

A second possibility is that the various properties that define a neuron class are subgrouped and may be independently invoked¹³. Thus the *unc-4* gene may be used to invoke the presynaptic specificity of VA neurons in certain contexts, whereas the directional preference for axon growth is invoked independently. This suggests that genes may exist that specify synaptic connectivity yet have no part in the mechanics of synaptogenesis. □

Received 3 October; accepted 26 November 1991.

1. Blochliger, K., Bodmer, R., Jack, J., Jan, L. Y. & Jan, Y. N. *Nature* **333**, 629–635 (1988).
2. Uemura, T., Shepherd, S., Ackerman, L., Jan, L. Y. & Jan, Y. N. *Cell* **58**, 349–360 (1989).
3. Way, J. C. & Chalfie, M. *Cell* **54**, 5–16 (1988).
4. Hedgecock, E. M., Culotti, J. G. & Hall, D. H. *Neuron* **2**, 61–85 (1990).
5. Elkins, T., Zinn, K., McAllister, L., Hoffman, F. M. & Goodman, C. S. *Cell* **60**, 565–575 (1990).
6. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. *Phil. Trans. R. Soc. B* **314**, 1–340 (1986).
7. Brenner, S. *Genetics* **77**, 71–94 (1974).
8. Wood, W. B. (ed.) *The Nematode Caenorhabditis elegans* (Cold Spring Harbor Laboratory Press, New York, 1988).

9. Chalfie, M. et al. *J. Neurosci.* **5**, 956–964 (1985).
10. Miller, D. M. et al. *Nature* **355**, 841–845 (1992).
11. Sulston, J. E. & Horvitz, H. R. *Dev. Biol.* **56**, 110–156 (1977).
12. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. *Dev. Biol.* **100**, 64–119 (1983).
13. Desai, C., Garriga, G., McIntyre, S. L. & Horvitz, H. R. *Nature* **336**, 638–646 (1988).
14. Durbin, R. M. thesis, Cambridge Univ. (1987).

ACKNOWLEDGEMENTS. We thank S. Hekimi, J. Hodgkin, D. Miller and G. Ruvkun for criticisms of the manuscript.

C. elegans *unc-4* gene encodes a homeodomain protein that determines the pattern of synaptic input to specific motor neurons

David M. Miller*, Michael M. Shen†‡, Caroline E. Shamu†‡, Thomas R. Bürglin§, Gary Ruvkun§, Michelle L. Dubois*, Medeva Ghee* & Laura Wilson*

* Department of Cell Biology, Duke University Medical Center, Durham, North Carolina 27710, USA

† MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK

§ Department of Genetics, Harvard Medical School, and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

THE creation of neural circuits depends on the formation of synapses between specific sets of neurons. Little is known, however, of the molecular mechanisms governing synaptic choice. A mutation in the *unc-4* gene alters the pattern of synaptic input to one class of motor neurons in the *Caenorhabditis elegans* ventral nerve cord. In *unc-4(e120)*, the presynaptic partners of VA motor neurons are replaced with interneurons appropriate to motor neurons of the VB class. This change in neural specificity is not accompanied by any detectable effects on neuronal morphology or process extension^{1,2}. We show that the absence of a functional *unc-4* gene product accounts for the mutant phenotype. The *unc-4* gene encodes a homeodomain protein and thus is likely to function as a transcription factor. The limited effect of the *unc-4* null mutation on cell fate may mean that *unc-4* regulates the expression of a small number of target genes and that the products of these genes are directly involved in the choice of synaptic partners.

We used a noncomplementation screen to isolate eight new *unc-4* alleles^{3,4} (Fig. 1a). One of these new alleles, *e2322ts*, was temperature-sensitive. A single spontaneous *unc-4* allele, *wd1*, was isolated from the transposon mutator strain TR679 (ref. 5)

in a separate screen for movement-defective animals. Evidence from our genetic studies indicates that *unc-4(e120)* represents a null allele. First, our noncomplementation screen failed to recover alleles with a more severe movement defect than *unc-4(e120)*, although it could recover genetic deficiencies for the *unc-4* locus. Indeed, seven of the eight alleles recovered in this screen possess a gross phenotype identical to that of *unc-4(e120)* at the dissecting-scope level (the exception being the temperature-sensitive allele *e2322ts*). Second, the movement defect of *e120* in *trans* to deficiencies of the locus is identical to that of *unc-4(e120)* homozygotes⁶. Third, the frequency of putative, null alleles recovered, about 1×10^{-4} , was similar whether using ethylmethanesulphonate (EMS) or ³²P as a mutagen, suggesting that the recovery of alleles was not dependent on the molecular nature of the mutation induced. Finally, five *unc-4* alleles are associated with deletions of a substantial fraction of the coding region of the gene. We conclude that the *e120* mutation effectively eliminates *unc-4* function and, thus, that the change in synaptic input to VA motor neurons that occurs in *unc-4(e120)* animals is the null phenotype.

Having demonstrated that synaptic input to VA motor neurons depends on an active *unc-4* gene, we used the *unc-4* temperature-sensitive allele, *unc-4(e2322ts)*, to determine if this requirement is restricted to a particular developmental period. VA motor neurons arise from the division of P-cell ectoblasts in the late L1 stage^{7,8}; the adult pattern of synaptic connectivity in the ventral nerve cord is established at the completion of the L1–L2 molt (J. White, personal communication). A temperature-shift experiment⁹ with *unc-4(e2322ts)* indicates that the creation of functional synapses between VA motor neurons and particular interneurons depends on *unc-4* activity around the time that these connections are made. The *unc-4* gene product is also necessary during the L2 and L3 stages, after the VA neurons have been wired into the ventral cord circuit at the end of the L1. These data suggest that *unc-4* may act to sustain the initial pattern of synaptic input to the VAs for several hours after it is first created. That embryos and young L1 larvae are not temperature-sensitive indicates that ventral cord wiring does not depend on *unc-4* function during the embryonic period in which interneurons providing input to the VAs are generated^{8,20}.

We cloned the *unc-4* gene first by identifying nearby Tc1 transposon insertion sites and then by using cosmid clones from this region of the *C. elegans* genome map¹⁰ to probe *unc-4* mutant DNA for restriction fragment length polymorphisms (RFLPs). From three factor crosses with flanking genetic

‡ Present addresses: Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA (M.M.S.); Department of Biochemistry, University of California at San Francisco, San Francisco, California 94143-0448, USA (C.E.S.).

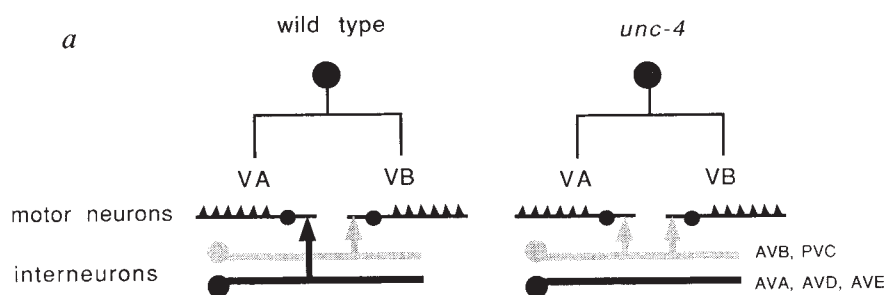
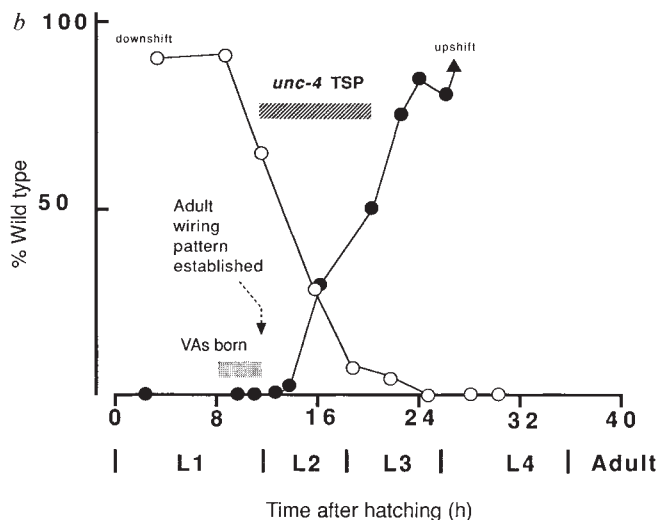


FIG. 1 *a*, Cell lineage and developmental fate of wild-type and *unc-4* motor neurons in *C. elegans* (adapted from ref. 1). During postembryonic development, a common precursor cell gives rise to most of the VA and VB motor neurons in the ventral nerve cord^{7,8}. VA and VB neurons adopt similar but mirror image morphologies. The elongated projection from each cell body depicts an axon and the solid triangles symbolize neuromuscular junctions. The shorter dendritic processes receive input (arrows) from different sets of interneurons; VBs accept input from interneurons AVB (gap junctions) and PVC (chemical synapses) whereas VA motor neurons are connected to interneurons AVA (gap junctions and chemical synapses), AVD and AVE (chemical synapses)². In *unc-4(e120)*, most of the VA motor neurons acquire input (gap junction) from AVB instead of the usual connections with AVA, AVD and AVE. As a result of the loss of appropriate input to the VAs, *unc-4(e120)* mutants are unable to crawl backwards. The change in wiring is limited to the presynaptic partners of the nine VA motor neurons (VA2–VA10) that are born with VB sisters. Synaptic input to the three VA motor neurons (VA1, VA11, VA12) that do not have VB sisters is not altered in *unc-4(e120)*. Therefore, *unc-4* activity is necessary to distinguish the fates of VA and VB neuroblasts that arise as sister cells but is not required to specify presynaptic input to VA motor neurons derived from other sub-lineages. It is important to note that the wiring defect in *unc-4(e120)* occurs in the absence of any detectable effect on either cell morphology or the position of neuronal processes in the ventral cord fascicle¹. *b*, *unc-4* temperature-sensitive period. The ability of the *unc-4* mutant *e2322ts* to crawl backwards is nearly wild type when these animals are reared at the permissive temperature of 16 °C but is 'Unc' (that is, unable to back up) when kept at the restrictive temperature of 25 °C. Synchronized cultures of *unc-4(e2322ts)* were shifted from 25 °C to 16 °C (○) or from 16 °C to 25 °C (●) at the indicated time points. Times are given in 25 °C equivalents. The closed triangle refers to *unc-4(e2322ts)* animals that were grown at 16 °C throughout. The movement phenotypes were scored later when animals were young adults. The temperature-sensitive period (TSP) is defined as the minimum time in which *unc-4* must be active to establish the wild-type pattern of wiring in the ventral nerve cord. As indicated, VA motor neurons arise during the late L1 stage (grey rectangle) and are incorporated into the ventral nerve cord circuitry by the end of the L1–L2 molt. Although the TSP shown in *b* does not overlap the late L1 stage during which VAs are generated and accept input from specific interneurons, this apparent discrepancy is within the resolution limits of this temperature shift experiment (see below and ref. 9). Thus, it is reasonable to assume that the *unc-4* gene product functions during this period and is required to select the presynaptic partners for the VAs in the late L1.

METHODS. *a*, We induced an allele of *unc-4* on the linkage group II balancer chromosome *mnC1*, to screen for new alleles conveniently⁴. By mutagenizing a stable, balanced strain containing this new allele as a heterozygote, *rol-6(e187) bli-1(e769)/mnC1(unc-4(e2151))*, and then by selecting for *Unc-4* F₁ progeny, we identified five new EMS-induced *unc-4* alleles (*e2307*, *e2320*, *e2321*, *e2322ts*, *e2323*) and three new ³²P-induced *unc-4* alleles (*e2308*,

markers³ we established that two of these TcI-containing fragments, *eP43* and *eP44*, are tightly linked to the *unc-4* locus (Fig. 2a). Cosmid clones covering the *eP43* and *eP44* insertion sites were obtained from the collection of overlapping cosmid and YAC clones, which now includes most of the *C. elegans* genome¹¹. These cosmids, ZK885 and ZK892, were used as probes to position *eP43* and *eP44* relative to various genetic deletions spanning the *unc-4* region. In some cases, presumptive deletion breakpoints were also identified as RFLPs on separate blots probed with other nearby cosmids. Taken together, the deletion and breakpoint mapping data defined a region of about 100 kilobases (kb) between *eP43* and *eP44* in which the *unc-4*



e2309, *e2314*) on the *rol-6(e187) bli-1(e769)*-marked chromosome. *b*, Worms were grown³ at either 16 °C or 25 °C and synchronized by allowing embryos to hatch at the starting temperature on a 20-μm nylon screen (Spectrum) dipped in a pool of M9 buffer. From 2 h (25 °C) to 3 h (16 °C) after hatching began, L1 larvae that crawled through the screen were distributed over several 60-mm NGM-OP50 plates pre-equilibrated to either 16 °C or 25 °C. Individual plates were transferred at intervals to a second incubator set to the alternative temperature. Developmental age at the time of temperature shift was confirmed by direct observation of a sample from each population by Nomarski optics⁷. Phenotypes were determined after 48 h or after the worms were young adults. Individual animals were tapped on the head with a platinum wire; worms that could sustain at least two cycles of contractile waves for backing up were scored as wild type as this is comparable to the prevailing phenotype of *unc-4(e2322ts)* animals grown at permissive temperature. At least 100 animals were scored for each time point. The boundaries of the TSP were established according to criteria suggested in⁹. The beginning of the TSP corresponds to the first point in the downshift experiment in which the fraction of phenotypically wild-type animals decreases; the end of the TSP corresponds to the last point in the upshift experiment at which there are detectably fewer wild-type animals than in a population maintained at permissive temperature throughout. By this criterion, the TSP could be expanded both to the left and to the right of the current boundaries to arbitrarily chosen timepoints just inside the apparent plateau percentages of wildtype animals but we have not done so because we do not have temperature shift data from these intervals.

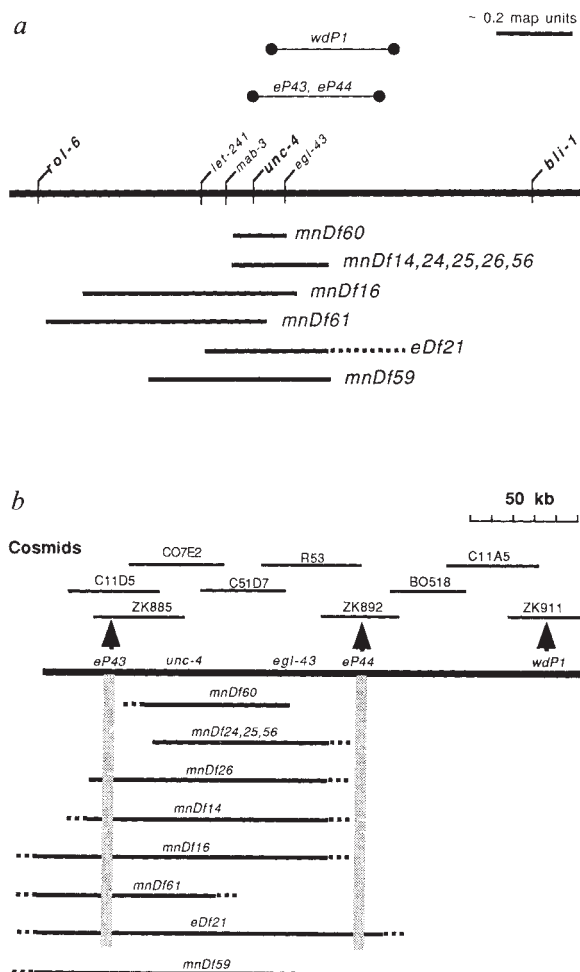
gene was likely to be located (Fig. 2b).

Next, we probed Southern blots of DNA from *unc-4* mutants with cosmids from the *eP43*–*eP44* region. One of these cosmids, CO7E2, detected deletions in five different *unc-4* alleles, *e887*, *e2308*, *e2320*, *e2314* and *wd1* (Fig. 3a). A 1.8-kb *Eco*R1 fragment is either altered or deleted in all five of these *unc-4* mutants. One of these alleles, *unc-4(e2320)*, corresponds to an internal deletion of about 200 base pairs (bp) from the 1.8-kb *Eco*R1 fragment (Fig. 3b). Molecular abnormalities were not observed in any of the other 12 *unc-4* alleles examined by this method (data not shown).

Restriction mapping CO7E2 showed that the 1.8-kb *Eco*R1

FIG. 2 Genetic and physical maps of *unc-4* region of chromosome II. *a*, Genetic map of the interval from *rol-6* to *bli-1*. The extents of deficiencies in *mnDf* strains are indicated as solid lines at the bottom⁶. The right end of *eDf21* has not been defined genetically and is depicted with a dashed line²⁴. Only those genetic markers and deficiency strains used in this study for breakpoint and polymorphism mapping are shown. The bars for the three Tc1 polymorphisms (*eP43*, *eP44*, *wdP1*) represent 95% confidence levels for their locations. The map distance for the *rol-6* to *bli-1* interval is 1.4 map units. *b*, Physical map of *unc-4* region. Cosmids from the region^{10,11} are shown at the top. The arrows point to specific cosmids that cover insertion sites for Tc1 polymorphisms, *eP43*, *eP44* and *wdP1*. Deletion endpoints in each deficiency strain are shown at the bottom as solid lines; dashed lines indicate deficiency breakpoints that have not been mapped to specific cosmids. Deficiencies that delete either the *eP43* or *eP44* Tc1 insertion sites intersect the vertical grey line. The scale bar of 50 kb is an estimate based on the average size of *HindIII* fragments used in establishing overlaps between cosmid clones¹⁰. The placement of *egl-43* over the right end of *mnDf60* is based on the genetic observation that *egl-43* is partially complemented by *mnDf60* (ref. 25).

METHODS. *a*, The mapping of Tc1 polymorphisms relies on the standard wild-type (N2) strain of *C. elegans* having a low-copy number of Tc1 elements (about 30), whereas other strains can have up to 500 elements, distributed throughout the genome²⁶. Using three factor crosses³, we constructed congenic strains in which the DNA derived from a high-copy-number strain, DH424 (ref. 27) or N62 (S. Brenner, personal communication) is limited to a small region around *unc-4*, between the flanking markers *rol-6* (roller) and *bli-1* (cuticular blisters). High-copy-number strain derived DNA polymorphisms *eP43* and *eP44* were identified as novel restriction fragments on Southern blots²⁸ probed with a cloned Tc1 element. The genetic map positions of these polymorphisms were obtained from Southern blot analysis with the Tc1 probe after recombination between the *rol-6 bli-1* marked congenic chromosomes and wild-type (N2) chromosomes marked with either *unc-4* or *let-241 unc-4*. In these experiments, *eP43* and *eP44* were genetically inseparable from *unc-4*. Using similar methods, the Tc1 polymorphism *wdP1* was closely linked to the *unc-4* allele *wd1* (~0.1 map units to the right) which arose spontaneously from the Tc1 mutator strain, TR679 (ref. 5). The Tc1-containing restriction fragments *eP43*, *eP44*, and *wdP1* were subcloned from agarose gel slices and unique DNA sequences from the Tc1-flanking regions used as probes to obtain overlapping genomic phage clones²⁸. These were matched with cosmid clones in the physical map of the *C. elegans* genome by A. Coulson and J. Sulston using standard techniques^{10,11}. *b*, To map *eP43* and *eP44* against genetic deletions in the region, polymorphism/deficiency (*eP/Df*) strains were constructed by placing either *rol-6 eP43 bli-1* or *rol-6 eP44 bli-1* chromosomes in *trans* to *Df* chromosomes spanning the *unc-4* region⁶. These heterozygous strains are effectively balanced because the *Df/Df* homozygotes are dead and *rol-6 bli-1* strains have small brood sizes and grow slowly. Southern blots of DNA from each *eP/Df* strain were probed with cosmids overlapping the *eP43* or *eP44* Tc1 insertion sites (that is, ZK885 and ZK892, respectively). If a given deficiency deletes a polymorphic site, only the polymorphic band arising from the *eP43* or *eP44* chromosome will be present on the blot. If a deficiency does not delete the polymorphic site, both a polymorphic band (from the *eP43* or *eP44* chromosome) and 'wild-type' band (from the chromosome bearing the deficiency) are present²⁹. The positions of physical breakpoints of genetic deficiencies were detected as novel hybridizing bands on Southern blots of *Df* strains probed with a series of individual cosmids from the region. The deficiency mapping data indicate that *unc-4* is flanked



by *eP44* on the left and *eP43* on the right: (1) The orientation of the physical map with respect to the 'right-left' convention of the genetic map was defined by the position of *wdP1* which genetic data positioned to the right of *unc-4*. (2) Neither *eP43* nor *eP44* are deleted by the smallest deficiency, *mnDf60*, which indicates that both polymorphisms are physically separate from *unc-4*. (3) *mnDf16* and *mnDf61* extend from just to the right of *unc-4* almost to *rol-6* (~0.7 map units) but neither deficiency deletes *eP44*; because our genetic map data indicate that *eP44* is to the right of *let-241*, which is covered by *mnDf16* and *mnDf61*, it follows that *eP44* must also be located to the right of these deficiencies and hence to the right of *unc-4*. (4) The *eP43* insertion site is deleted by several deficiencies including *mnDf61*. If *eP43* were located to the right of *unc-4*, then its insertion site should also be deleted by *mnDf24*, *mnDf25* and *mnDf56* which cover the right end of *mnDf61* but which do not also cover *eP43*. The leftward position of *eP43* with respect to *unc-4*, however, does fit these data.

fragment is near a previously identified open reading frame, *ceh-4*, encoding the carboxyl terminal 14 amino acids of a homeobox motif¹² (Fig. 3c). We sequenced the 1.8-kb *EcoRI* fragment and identified two potential exons encoding amino-acid sequences homologous to the first 47 residues of a canonical homeodomain that splice to the remaining 14 homeodomain amino acids defined by *ceh-4*. Thus, our genomic sequence data agree with a model in which a 61 amino-acid homeodomain¹³ is translated from a spliced transcript derived from three exons. Two other exons were identified in the sequence of the genomic 1.8-, 0.5- and 1.05-kb *EcoRI* fragments. The DNA sequence of polymerase chain reaction (PCR)-derived complementary DNA clones confirmed this interpretation of the genomic sequence data (Fig. 3c).

Northern blots probed with the 1.8-kb *EcoRI* fragment revealed a low abundance 1.2-kb transcript in wild-type (N2) poly(A)⁺ RNA (Fig. 3d) that was not present in poly(A)⁺ RNA

isolated from either *unc-4(wd1)* or from *unc-4(e2308)* (data not shown). Because no other RNA species were detected by the 1.8-kb probe, it is reasonable to conclude that the 1.2-kb transcript is derived from the open reading frames that we have identified and that these exons represent *unc-4* coding sequence. Despite extensive screening of cDNA libraries with the 1.8-kb probe, no *unc-4* cDNA clones were detected. On the basis of the 1.2-kb length of the *unc-4* transcript, the PCR-derived cDNA clones include about half (552 bp) of the transcribed sequence (Fig. 4a).

Although the *unc-4* homeodomain is related to the *prd* family of homeobox motifs^{14,15}, it also has differences that may indicate a new type of homeodomain sequence (Fig. 4b). How does the *unc-4* gene control the pattern of synaptic input to VA motor neurons? Our evidence indicates that *unc-4* is likely to control neural specificity by regulating transcription of other genes¹⁶. It is unlikely, however, that *unc-4* regulates genes

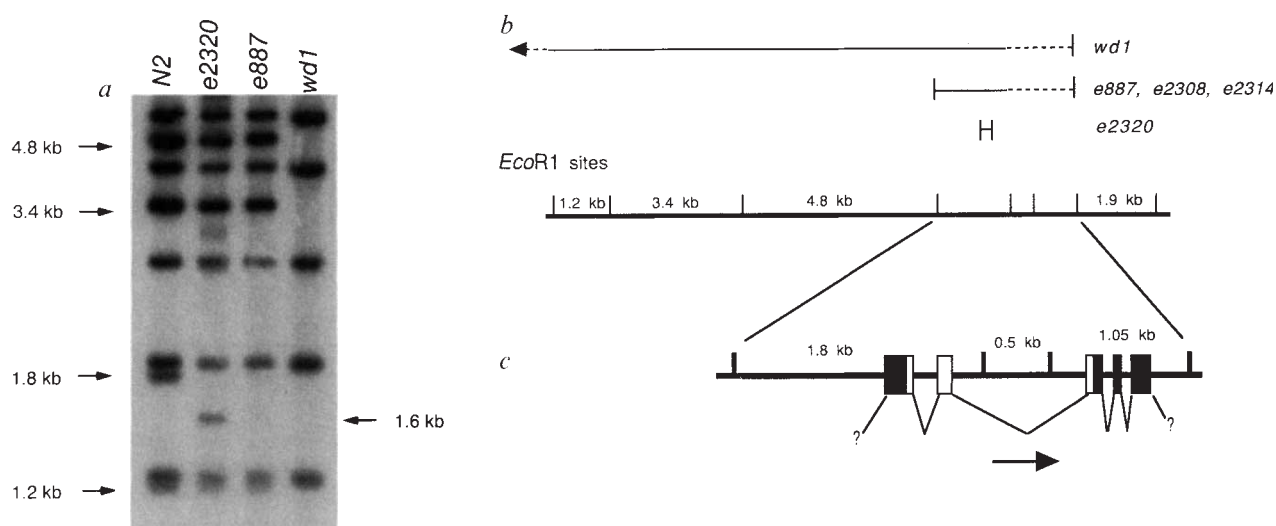


FIG. 3 *a*, Southern blot of *unc-4* allele-specific polymorphisms. Blots were prepared from *Eco*R1-digested wild-type (N2) and *unc-4* mutant DNA and probed with 32 P-labelled cosmid C07E2. DNA from all 17 of the known *unc-4* alleles was examined but only e2320, e887 and wd1 are shown. Molecular size labels on the left refer to *Eco*R1 fragments missing from *unc-4* mutants. In *unc-4*(e2320), a novel 1.6-kb *Eco*R1 fragment is present. *b*, Partial restriction map of *unc-4* region showing *Eco*R1 sites and rough positions of deletions in five different *unc-4* alleles. *unc-4*(e2320) corresponds to a 0.2-kb internal deletion of the 1.8-kb *Eco*R1 fragment. In wd1, e887, e2308 and e2314, the 1.8-kb *Eco*R1 fragment is absent. Additional polymorphic bands were detected in *unc-4*(e2314) which may correspond to a molecular rearrangement in a nearby region. Dashed lines are indicative of deletion endpoints that have not been mapped relative to the underlying *Eco*R1 restriction sites. *c*, Exons in the genomic region affected by *unc-4* mutations. Exons are indicated by the boxes. The proposed intron-exon boundaries were identified from conserved splice junction sequences³⁰ and confirmed by sequencing PCR-derived cDNA clones with the exception of the 5' end of the first exon and the 3' end of the last exon for which we do not have overlapping cDNA sequence data (see Fig. 4). The homeobox sequence corresponds to the unshaded portions of exons. The direction of transcription is indicated by the arrow. *d*, Northern blot of wild-type and *unc-4* mutant RNA. A northern blot of polyadenylated (poly(A⁺)) RNA from wild-type (N2) and *unc-4* (wd1) was hybridized with a 32 P-labelled antisense riboprobe from the 1.8-kb *Eco*R1 fragment. A 1.2-kb band is detected in N2 poly(A⁺) RNA but not in RNA from *unc-4*(wd1). The blot was separately hybridized with an actin riboprobe to verify that equal amounts of RNA were loaded in each lane. Similar results were obtained with *unc-4*(e2308) (data not shown). METHODS. *a*, Genomic *C. elegans* DNA³¹ from the 17 known *unc-4* mutant strains was digested with *Eco*R1, Southern blotted, and probed with 32 P-labelled cosmid C07E2 (ref. 28). The *unc-4* alleles examined were e26, e120, e521, e581 (ref. 3), e887 (ref. 32), H7*3 (G. Garriga, personal communication), e2151, e2152, e2307, e2308, e2309, e2314, e2320, e2321, e2322ts, e2323 and wd1 (this work). Separate hybridization experiments with the 1.8-kb *Eco*R1 fragment showed that the novel 1.6-kb band in *unc-4*(e2320) DNA is derived from the 1.8-kb wild-type *Eco*R1 fragment (data not shown). *c*, The 1.8-kb, 0.5-kb and 1.05-kb *Eco*R1 fragments were subcloned into Bluescript and sequenced using a combination of *Exo*III deletions and specific oligonucleotide primers²⁸. *d*, RNA was isolated as described in ref. 33, subjected to additional phenol-chloroform extractions

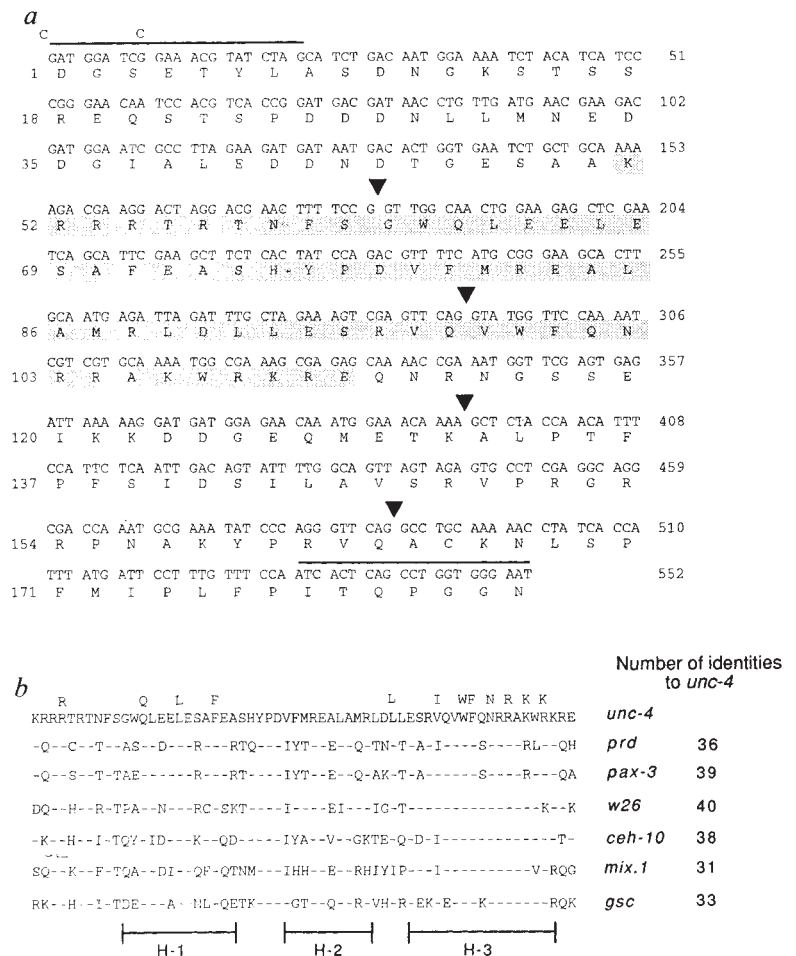
required for axonal outgrowth or guidance because the elimination of *unc-4* activity has no detectable effect on process morphology or placement in the ventral cord fascicle. Nor does the establishment of functional synapses depend on *unc-4* activity because these are created in *unc-4* mutants, albeit with abnormal sets of interneurons¹. The limited effect of the *unc-4* mutation on neuronal fate (Fig. 1a), may mean that *unc-4* controls relatively few target genes and that the products of these genes are directly involved in the choice of synaptic partners.

In this respect, *unc-4* is unique among homeodomain genes

followed by ethanol precipitation to improve stability, and treated with RNase-free DNase (Boehringer Mannheim). Six μ g total RNA and 6 μ g poly(A⁺) RNA for each nematode strain were used for northern blots²⁸. 32 P-labelled RNA probes were prepared with a T7-T3 riboprobe kit and northern blots hybridized according to the recommendations of the manufacturer (Stratagene). Filters were washed at a final stringency of 0.1 $\times$ SSC, 0.1% SDS for 30 min at 75 $^{\circ}$ C. Transcript size was estimated by comparison to an RNA molecular mass standard (BRL). Blots were first hybridized with the 1.8-kb *Eco*R1 antisense riboprobe and exposed to X-ray film for 40 h with an intensifying screen at -80 $^{\circ}$ C. The blots were stripped by boiling in 0.5% SDS and then hybridized with an actin antisense riboprobe (p17/13 18-103)³⁴. Autoradiography of blots hybridized with the actin probe was for 1 h.

that function in neurogenesis. Mutations in other homeodomain genes are accompanied by transformations of one neuronal type to another¹⁷⁻¹⁹. For example, in *Drosophila*, the elimination of *ftz* or *eve* activity causes RP2 neurons to adopt the pattern of axonal outgrowth and fasciculation normally assumed by their sister RP1 neurons^{17,18}. In these cases, neural wiring defects can be attributed to substantial changes in cell fate that are likely to result from aberrant expression of several classes of gene products. Thus, *unc-4* may define a new category of homeodomain gene which functions during the final stages of neural

FIG. 4 a, Partial *unc-4* nucleotide coding sequence and amino acid sequence of presumptive *unc-4* gene product. PCR-derived cDNA sequences were checked against genomic sequence data from overlapping regions of the 1.8-kb and 1.05-kb *EcoRI* fragments shown in Fig. 3c. PCR primers used to generate cDNA are denoted with solid black lines. The homeobox is outlined by the shaded box. Triangles show the positions of introns in the genomic sequence. The region flanking the amino-terminal end of the homeodomain is rich in acidic amino acid sidechains ($pI \approx 3.2$) and hydrophilic but is not closely related to comparable regions of other homeodomain-containing proteins. **b**, Comparison of homeodomains. The deduced amino-acid sequence of the *unc-4* homeodomain (single-letter code) is aligned with representative members of the *paired* family of homeodomains, *prd* (ref. 14) and *pax-3* (ref. 15) and with *prd*-related homeodomain sequences, *ceh-10* (ref. 35), *w26* (U. Waldorf and W. Gehring, personal communication), *mix.1* (ref. 36) and *gooseoid* (*gsc*) (ref. 37). Dashes indicate amino acids that are shared with *unc-4*. Twelve amino acids that are highly conserved in all homeodomains are listed across the top^{13,16}. The helical regions of the homeodomain structure, helix-1 (H-1), helix-2 (H-2) and helix-3 (H-3) are shown at the bottom³⁸. Compared with the known families of homeodomain sequences^{13,16}, the *unc-4* homeodomain is most closely related to the *prd* family (55–65% identity) but this similarity is substantially lower than the degree of homology between the *Drosophila* and mouse *prd*-type homeodomains which are 80–85% identical¹⁵. The most significant difference occurs in the third helical domain or DNA recognition helix. Studies of homeodomain-DNA interaction have established that binding specificity is critically dependent on the ninth amino acid of helix-3 (residue 51) which is glutamine in *unc-4* and serine in the *prd*-type genes³⁹. In addition, the *Drosophila* and mouse *prd* family genes include three regions of homology outside of the homeodomain that we have not detected in the *unc-4* sequence¹⁵. The *unc-4* homeodomain joins a growing group of homeobox motifs (*ceh-10*, *w26*, *mix.1*, *gsc* and others) which are related to the *prd* family but which also exhibit important differences from the consensus *prd* sequence. Of this set of *prd*-related homeodomains, the homeobox sequence of the *Drosophila* gene, *w26*, with 40/61 matches, is most closely related to the *unc-4* homeobox. The homology is highest near the carboxyl end of the homeodomain where the DNA recognition helical domains of *w26* and *unc-4* are virtually identical. The *C. elegans* homeodomain most closely related to *unc-4* is from the *ceh-10* gene (38/61 matches)³⁵. **METHODS**. A 552-bp cDNA fragment was generated by reverse transcription-



PCR from total RNA⁴⁰ using the indicated pair of oligonucleotide primers (solid lines). A single base mismatch (G to C) was introduced into the 5' primer to create a unique *Bam*H1 site for subcloning. The cDNA sequence confirmed the existence of the five exons previously identified by inspection of genomic DNA sequences. MacVector 3.5 was used for sequence analysis and for searching Genbank for sequence homologies.

differentiation and which regulates a much smaller group of downstream loci.

Other homeobox genes that regulate neurogenesis in *C. elegans* are expressed in the cells whose developmental fate depends on them^{21,22}. It is also possible that *unc-4* is activated

in a nearby cell which in turn provides an inductive signal²³ to the VA neuron or its presynaptic partners. *In situ* localization of the *unc-4* protein with specific antibodies²² or fusing the *unc-4* promoter to a β -galactosidase reporter gene²¹ should distinguish between these alternatives. □

Received 11 September; accepted 12 December 1991.

- White, J. G., Southgate, E. & Thomson, J. N. *Nature* **355**, 838–841 (1992).
- White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. *Phil. Trans. R. Soc. B* **314**, 1–340 (1986).
- Brenner, S. *Genetics* **77**, 71–94 (1974).
- Herman, R. K. *Genetics* **88**, 49–65 (1978).
- Collins, J., Saari, B. & Anderson, P. *Nature* **328**, 726–728 (1987).
- Sigurdson, D. C., Spanier, G. J. & Herman, R. K. *Genetics* **108**, 331–345 (1984).
- Sulston, J. E. & Horvitz, H. R. *Dev. Biol.* **56**, 110–116 (1977).
- Sulston, J. E. *Cold Spring Harb. Symp. quant. Biol.* **48**, 443–452 (1983).
- Suzuki, D. *Science* **170**, 695–706 (1970).
- Coulson, A., Sulston, J., Brenner, S. & Karn, J. *Proc. natn. Acad. Sci. U.S.A.* **83**, 7821–7825 (1986).
- Coulson, A. *et al. Nature* **355**, 184–186 (1990).
- Bürglin, T. R., Finney, M., Coulson, A. & Ruvkun, G. *Nature* **341**, 239–243 (1989).
- Scott, M. P., Tamkun, J. W. & Hartzell, G. W. *III BBA Rev. Cancer* **989**, 25–48 (1989).
- Bopp, D., Burri, M., Baumgartner, S., Frigerio, G. & Noll, M. *Cell* **47**, 1033–1040 (1986).
- Goulding, M. D., Chalepakis, G., Deutsch, U., Erselius, J. R. & Gruss, P. *EMBO J.* **10**, 1135–1147 (1991).
- Hayashi, S. & Scott, M. P. *Cell* **63**, 883–894 (1990).
- Doe, C. Q., Hiromi, Y., Gehring, W. J. & Goodman, C. S. *Science* **239**, 170–175 (1988).
- Doe, C. Q., Smouse, D. & Goodman, C. S. *Nature* **333**, 376–378 (1988).
- Jan, Y. N. & Jan, L. Y. *Trends Neurosci.* **13**, 493–498 (1990).
- Durbin, R. M. thesis, Cambridge Univ. (1987).
- Way, J. C. & Chaffie, M. *Genes Dev.* **3**, 1823–1833 (1989).
- Finney, M. & Ruvkun, G. *Cell* **63**, 895–905 (1990).
- Tomlinson, A., Kimmel, B. E. & Rubin, G. M. *Cell* **55**, 771–784 (1988).
- Shen, M. M. & Hodgkin, J. *Cell* **54**, 1019–1031 (1988).
- Shen, M. M. thesis, Cambridge Univ. (1988).
- Emmons, S. W., Yesner, L., Ruan, K. & Katzenberg, D. *Cell* **32**, 55–65 (1983).

- Liao, L. W., Rosenzweig, B. & Hirsh, D. *Proc. natn. Acad. Sci. U.S.A.* **80**, 3585–3589 (1983).
- Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning, a Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989).
- Shamu, C. E. thesis, Cambridge Univ. (1989).
- Emmons, S. W. in *The Nematode Caenorhabditis elegans* (ed. Wood, W. B.) 47–79 (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1988).
- Sulston, J. E. & Hodgkin, J. in *The Nematode Caenorhabditis elegans* (ed. Wood, W. S.) 587–606 (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1988).
- Babu, P. & Brenner, S. *Mutat. Res.* **82**, 269–273 (1981).
- Austin, J. & Kimble, J. *Cell* **58**, 565–571 (1989).
- Krause, M., Wild, M., Rosenzweig, B. & Hirsh, D. *J. Molec. Biol.* **208**, 381–392 (1989).
- Hawkins, N. C. & McGhee, J. D. *Nucleic Acids Res.* **18**, 6101–6106 (1990).
- Rosa, F. M. *Cell* **57**, 965–974 (1989).
- Blumberg, B., Wright, C. V. E., DeRobertis, E. M. & Cho, K. W. Y. *Science* **253**, 194–196 (1991).
- Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. T. & Pabo, C. O. *Cell* **63**, 579–590 (1990).
- Hanes, S. D. & Brent, R. *Science* **251**, 426–430 (1991).
- Kawasaki, E. S. in *PCR Protocols* (eds Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J.) 21–27 (Academic, New York, 1990).

ACKNOWLEDGEMENTS. We thank J. White for communicating his results to us before publication, A. Coulson and J. Sulston for providing cosmid and YAC clones, W. Gehring and U. Waldorf for the *w26* sequence, M. Krause for the pT7/T3 18–103 actin probe, J. Hodgkin and I. Greenwald for advice, T. O'Brien for comments on the manuscript and C. Niemeyer for technical assistance. D.M.M. thanks F. Schachar and M. Sheetz for their support. Some of the strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH NCRR. This work was supported by the NIH (D.M.M.), by graduate fellowships from the NSF (C.E.S. and M.M.S.) and by EMBO and Swiss National FONDS Fellowships (T.R.B.).

Clonal expansion of p53 mutant cells is associated with brain tumour progression

David Sidransky*, Tom Mikkelsen†, Karl Schwachheimer‡, Mark L. Rosenblum§, Web Cavanee† & Bert Vogelstein*||

* The Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA

† Ludwig Institute for Cancer Research, Montreal, Province of Québec, H3A 1A1, Canada

‡ Abteilung Neuropathologie Pathologischer Institut, Albert Ludwigs Universität, Freiburg, Germany

§ Brain Tumor Research Center, Department of Neurological Surgery, University of California Medical Center, San Francisco, California 94143, USA

TUMOUR progression is a fundamental feature of the biology of cancer¹. Cancers do not arise *de novo* in their final form, but begin as small, indolent growths, which gradually acquire characteristics associated with malignancy. In the brain, for example, low-grade tumours (astrocytomas) evolve into faster growing, more dysplastic and invasive high-grade tumours (glioblastomas)^{2,3}. To define the genetic events underlying brain tumour progression, we analysed the p53 gene in ten primary brain tumour pairs. Seven pairs consisted of tumours that were high grade both at presentation and recurrence (group A) and three pairs consisted of low-grade tumours that had progressed to higher grade tumours (group B). In group A pairs, four of the recurrent tumours contained a p53 gene mutation; in three of them, the same mutation was found in the primary tumour. In group B pairs, progression to high grade was associated with a p53 gene mutation. A subpopulation of cells were present in the low-grade tumours that contained the same p53 gene mutation predominant in the cells of the recurrent tumours that had progressed to glioblastoma. Thus, the histological progression of brain tumours was associated with a clonal expansion of cells that had previously acquired a mutation in the p53 gene, endowing them with a selective growth advantage. These experimental observations strongly support Nowell's clonal evolution model of tumour progression⁴.

Brain tumours, like most other tumour types, are associated with several genetic changes. Among these, loss or mitotic recombination of chromosome 17p is common⁵⁻⁷. Furthermore, in several glioblastomas⁸, as in many other tumour types^{9,10}, loss of one 17p allele correlates with mutation of p53 in the remaining allele. To evaluate the relationship between p53 and astrocytoma progression, tumour pairs were obtained from patients with brain tumours that had been surgically removed. Tumours were carefully dissected to ensure optimal removal of contaminating normal tissue. A portion of the p53 gene, including exons 5-8, was amplified by the polymerase chain reaction (PCR) and subcloned into a phagemid vector; pooled clones were sequenced as previously described¹¹. Four group A tumours contained p53 mutations at recurrence and, in three, the identical p53 mutation was seen at initial presentation (Table 1). All were missense mutations resulting in nonconservative amino-acid changes. From the ratio of wild-type to mutant p53 alleles in the sequencing gels, it was clear that all group A tumours with p53 gene mutations had lost the wild-type allele.

All three of the advanced tumours from group B (Fig. 1) had a p53 gene mutation. Again, all mutations were missense mutations resulting in nonconservative amino-acid changes (Table 1). Two of the three high-grade tumours (from patients B8 and B9) had lost the wild-type allele. No evidence of these mutations was observed in the less advanced tumours of the pairs in the sequencing assays. But were these mutations present in a small

fraction of cells in the initial tumour, in a proportion not detectable by sequencing? We used a sensitive plaque assay to identify a potential cell minority containing p53 mutations among a majority of cells with normal p53 genes. The p53 gene was amplified through PCR as described above and the PCR products subcloned into a phage vector. Specific oligonucleotides that recognize a mutant base pair in the p53 gene were then used to probe filter lifts of the phage plaques.

Specific oligonucleotides were synthesized that would detect the mutations seen in the tumours of patients B8 and B9 and used to probe the cloned PCR products of the low grade tumours. A small number of positive plaques were seen in the low-grade tumours (Fig. 2). Positive plaques were picked and the phage DNA sequenced to confirm the presence of the mutation. Blots were then stripped and hybridized with a probe detecting all p53 clones (mutant plus wild type) to calculate the percentage of mutant p53 genes in the low-grade tumours. This showed that 8% and 21% of the tumour cells from patients B8 and B9, respectively, contained a mutant p53 gene. The high-grade tumours from both patients consisted of cells in which over 95% contained mutant p53 genes as assessed by the same assay. Histological sections were examined and all tumours were found to be contaminated by <10% normal cells (Fig. 1).

The low-grade component of the tumour from patient B10 contained a p53 mutation at codon 253 and a light mutant band at codon 273 when pooled clones were sequenced, whereas the high-grade component was found to contain mutant bands of equal intensity at codons 253 and 273 (Table 1). After subcloning

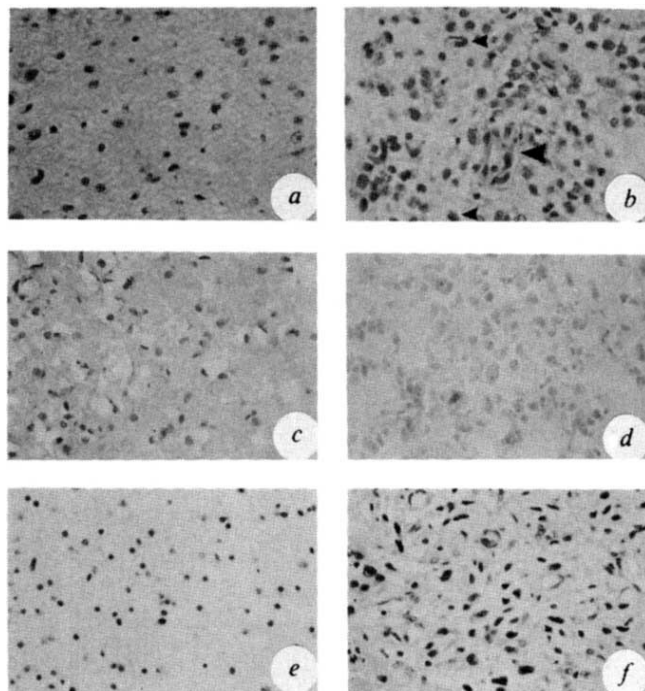


FIG. 1 Micrographs of tumours from patients in group B. The tumour from patient 10 had two distinct patterns (a and b). Most of the neoplasm was moderately cellular with a dense feltwork of glial processes. The nuclei of the glioma cells in a were mildly to moderately pleomorphic. Mitotic figures were rare and no vascular endothelial proliferation was identified. This region was classified as a moderately anaplastic astrocytoma. Sharply demarcated from this region and invading the leptomeninges was a more cellular component of the neoplasm with larger nuclei, nuclear pleomorphism and multiple mitotic figures (b, small arrows). The glial feltwork was less extensive and vascular endothelial proliferation (large arrow) was evident. This latter region was classified as glioblastoma multiforme. Both regions were glial fibrillary acid protein (GFAP) immunopositive, an indicator of their glial origin. The tumours in c and e also had features of low-grade glial tumours, whereas those in d and f had features of glioblastoma multiforme. a, c, and e, Initial low-grade tumours from patients B8, B9, and B10, respectively; b, d, and f, high-grade tumours from the same patients.

|| To whom correspondence should be addressed.

TABLE 1 Mutations in each tumour

Patient	Initial tumour grade	Initial tumour mutation*	Time to progression (months)	Progressed tumour grade	Progressed tumour mutation
Group A					
B1	IV	Codon 173 GTG → ATG, Val → Met	7	IV	Codon 173 GTG → ATG, Val → Met
B2	IV	Codon 244 GGC → AGC, Gly → Ser	13	IV	Codon 244 GGC → AGC, Gly → Ser
B3	IV	Codon 248 CGG → TGG, Arg → Trp	9	IV	Codon 248 CGG → TGG, Arg → Trp
B4	IV	None	11	IV	Codon 281 GAG → GAA, Asp → Glu
B5	IV	None	5	IV	None
B6	IV	None	9	IV	None
B7	IV	None	7	IV	None
Group B					
B8	II	None	15	IV	Codon 273 CGT → TGT, Arg → Cys
B9	I	None	28	IV	Codon 270 TTT → Ctt, Phe → Leu
B10	II	Codon 253 ACC → ATC, Thr → Leu	0	IV	Codon 253 ACC → ATC, Thr → Leu Codon 273 CGT → TGT, Arg → Cys

Group A tumours consisted of glioblastoma at both presentation and recurrence, whereas group B tumours were low-grade (I and II) at presentation and high-grade (III and IV) at recurrence. DNA was isolated from tumours and exons 5–8 of the p53 gene were amplified by the PCR, subcloned into lambda Zap, and sequenced as described previously¹¹. The tumour from patient B10 contained two components, histologically and mechanically separable, from the same tumour.

* As assessed by sequencing (see text).

amplified DNA, individual clones were sequenced and these two mutations were found to reside on separate alleles. Using the oligonucleotide-specific hybridization method described,

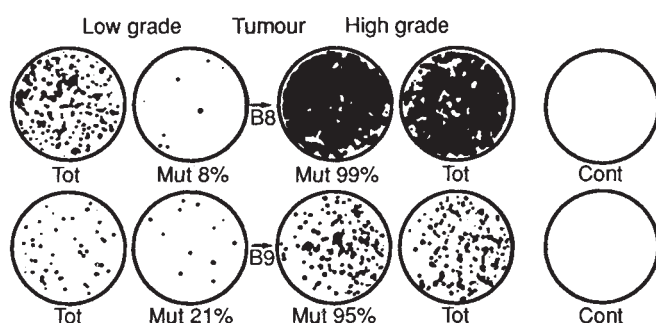


FIG. 2 Clonal expansion of p53 mutant cells. Over 95% of the clones obtained by PCR from low-grade and high-grade tumours in patients B8 (top) and B9 (bottom) hybridized to an oligonucleotide probe detecting all p53 clones (Tot). An oligonucleotide probe specific for mutant p53 at codon 273 hybridized to only a small number of plaques (8%) in the low-grade tumour from patient B8 (top left), but hybridized to the majority of plaques (99%) in the high-grade tumour. A mutant-specific oligonucleotide probe for the high-grade tumour in B9 also hybridized to 21% of clones in the low-grade tumour (bottom left) and 95% in the high-grade tumour. Neither of the mutant-specific probes hybridized to control plaques (far right) from another glioblastoma which contained a p53 mutation at a different codon.

METHODS. Tumour DNAs were amplified using PCR, ligated to lambda Zap arms (Stratagene), packaged, plated at a density of 200–500 plaques per plate, and the plaque DNAs transferred to nylon membranes as described¹¹. Mutant-specific oligomers corresponding to the mutant sequences in the high-grade tumours were synthesized to recognize a mutation at codon 273 (5'-TTGAGGTGTGTTTGTG-3') for tumours of patient B8 and B10; a mutation at codon 270 (5'-GGAACAGCCTTGAGGTGC-3') for tumours in patient B9; and a mutation at codon 253 (5'-CATCCTCATCATCAC-3') for the tumour in patient B10. As a control, the filters were rehybridized to an oligonucleotide probe for the wild-type sequence at codons 243 to 248 (5'-ATGGCGGCATGAACCGG-3'), which would identify all p53 clones (mutant or wild type, Tot). The oligomers were labelled with [³²P]ATP and T4 kinase to a specific activity of 5 × 10⁸ d.p.m. Hybridization with oligomers was at 45 °C for 1 h. After hybridization, the membranes were rinsed in 3 × SSC (450 mM NaCl, 18 mM sodium citrate, 1 mM Tris, pH 7.2), 0.1% SDS at room temperature for 5 min and then washed at 2–5 °C below the calculated melting temperature. The blots were then briefly dried and exposed to film. For the one-allele tumours from patients B8 and B9, the percentage of mutant cells was calculated by counting the mutant clones and dividing this number by the total number of clones that contained p53 sequences. We assumed that the subset of cells in the low-grade tumours which contained a p53 gene mutation also had a concomitant deletion of the wild-type allele, as was found in the high-grade tumours of these patients.

phage plaques from low- and high-grade components were probed with an oligonucleotide specific for the codon-253 mutation, an oligonucleotide specific for the 273 mutation, and a normal oligonucleotide detecting all p53 clones (mutant and wild type). This analysis revealed that 60% of the cells in the low-grade component contained one allele which was wild type and one allele that was mutant at codon 253, whereas both alleles were mutant (one at codon 253 and one at codon 273) in 40% of the cells. Virtually all of the cells in the high-grade component contained two mutant alleles. Thus progression appeared to be associated with mutation of the second allele at codon 273 in a cell with a pre-existing mutation in the first allele.

Nowell postulated that a cell acquiring a genetic change might acquire a selective growth advantage⁴. Clonal expansion of this cell, driven by successive mutations, would lead to tumour progression^{4,12}. Although the occurrence of several mutations has been observed in human tumours^{13,14}, evidence for clonal expansion of a subpopulation of cells with a specific endogenous mutation has until now not been available. The data presented here suggest that mutation of p53 leads to a selective growth advantage *in vivo* that seems to be a critical step in transformation from low-grade to high-grade tumours. This provides a direct experimental demonstration of clonal expansion in a human tumour: a rare cell carrying a specific change in a critical gene became the dominant cell type as the tumour progressed. This concept will probably be useful in understanding the role of various oncogenes and suppressor gene mutations in the pathway to malignancy. □

Received 8 November; accepted 12 December 1991.

1. Foulds, L. *Neoplastic Development* Vols 1 and 2 (Academic, London, 1969 and 1975).
2. Scherer, H. J. *Amer. J. Cancer* **40**, 159–198 (1940).
3. Russell, D. S. & Rubenstein, L. J. in *Pathology of Tumors of the Nervous System* 4th edn 226–228 (Williams and Wilkins, Baltimore, 1977).
4. Nowell, P. *Science* **194**, 23–28 (1976).
5. James, C. D. *et al. Cancer Res.* **48**, 5546–5551 (1988).
6. Bigner, S. H. *et al. Cancer Res.* **48**, 405–411 (1988).
7. James, C. D. *et al. Proc. natn. Acad. Sci. U.S.A.* **86**, 2858–2862 (1989).
8. Nigro, J. M. *et al. Nature* **342**, 705–708 (1989).
9. Baker, S. J. *et al. Science* **244**, 217–221 (1989).
10. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. *Science* **253**, 49–53 (1991).
11. Sidransky, D. *et al. Science* **252**, 706–709 (1991).
12. Kerbel, R. S. *Adv. Cancer Res.* **55**, 87–132 (1990).
13. Bishop, J. M. *Cell* **64**, 235–248 (1991).
14. Fearon, E. R. & Vogelstein, B. *Cell* **61**, 759–767 (1990).

ACKNOWLEDGEMENTS. We thank S. J. Armond for photomicrographs and pathological description of the tumour in patient B10, J. T. Vogelstein for reading, and T. Gwiazda for preparation of the manuscript. Supported by the Preuss Foundation and the NIH.

Genetically modified photosynthetic antenna complexes with blueshifted absorbance bands

G. J. S. Fowler, R. W. Visschers*, G. G. Grief, R. van Grondelle* & C. N. Hunter

Krebs Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield, S10 2UH, UK

* Department of Physics and Astronomy, Free University of Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands

LIGHT energy for photosynthesis is collected by the antenna system, creating an excited state which migrates energetically 'downhill'. To achieve efficient migration of energy the antenna is populated with a series of pigments absorbing at progressively redshifted wavelengths. This variety in absorbing species *in vivo* has been created in a biosynthetically economical fashion by modulating the absorbance behaviour of one kind of (bacterio)chlorophyll molecule. This modulation is poorly understood but has been ascribed to pigment-pigment and pigment-protein interactions. We have examined the relationship between aromatic residues in antenna polypeptides and pigment absorption, by studying the effects of site-directed mutagenesis on a bacterial antenna complex. A clear correlation was observed between the absorbance of bacteriochlorophyll *a* and the presence of two tyrosine residues, α Tyr44 and α Tyr45, in the α subunit of the peripheral light-harvesting complex of *Rhodobacter sphaeroides*, a purple photosynthetic bacterium that provides a well characterized system for site-specific mutagenesis¹⁻³. By constructing single (α Tyr44, α Tyr45 $\rightarrow$ PheTyr) and then double (α Tyr44, α Tyr45 $\rightarrow$ PheLeu) site-specific mutants, the absorbance of bacteriochlorophyll was blueshifted by 11 and 24 nm at 77 K, respectively. The results suggest that there is a close approach of tyrosine residues to bacteriochlorophyll, and that this proximity may promote redshifts *in vivo*.

Site-specific mutagenesis of plasmid-borne DNA requires that the parental genomic sequences for LH2 are deleted. Accordingly, genes encoding polypeptides for the LH2 antenna complex were deleted from the chromosome and replaced with an antibiotic resistance cassette; this created an LH2⁻LH1⁺RC⁺ recipient. As a further modification, the genes encoding the LH1 and reaction centre complexes were also deleted and replaced with a kanamycin-resistance cassette; this resulted in a recipient strain devoid of any photosynthetic complexes (that is, LH2⁻LH1⁻RC⁻). Site-directed mutagenesis was used to change α Tyr44 and α Tyr45 to Phe and Leu, respectively (Fig. 1), thereby creating an LH2 complex that resembles B800-820 in *Rhodospirillum rubrum*.

In Fig. 2, the 77 K absorbance spectra of plasmid-borne, wild-type LH2, the single mutant α Tyr44 $\rightarrow$ Phe and the double mutant α Tyr44 $\rightarrow$ Phe, α Tyr45 $\rightarrow$ Leu are shown. It is clear that when the wild-type gene is reintroduced, a spectrum is observed that is very similar to the absorbance spectrum of purified B800-850 complexes⁴. The two bands have maximal absorbance at 850 and 798 nm and the intensity ratio of the two peaks is 0.8. In the single α Tyr44 $\rightarrow$ Phe mutant the absorbance maximum of the 850 nm band is shifted 11 nm to 839 nm. A slight decrease in the intensity of this band relative to the 800 nm peak is also observed. In the double mutant α Tyr44 $\rightarrow$ Phe, α Tyr45 $\rightarrow$ Leu we observed a further shift of the 850 nm band to 826 nm, and the intensity of this band relative to the 800 nm peak is further decreased; the explanation for this reduction in intensity is unknown. The absorbance spectrum of membranes from this mutant is very similar to the spectrum of purified B800-820 complexes⁵.

Excitation spectra at 77 K reveal some of the energy transfer processes occurring in these antennae, and demonstrate that the 800 nm bacteriochlorophyll is fully able to transfer energy to, and elicit fluorescence from, the longer-wavelength components (Fig. 3a). Comparison of Fig. 3 with the absorbance spectra in Fig. 2 demonstrates that all of the altered LH2 participates in energy transfer to LH1 because there were no significant changes in the intensity ratios of the peaks.

Circular dichroism (CD) has been used to provide information on light-harvesting complexes because it is a sensitive probe of the associations between pigments in this type of system⁶. Native LH2 shows a strong conservative CD centred at 850 nm that may arise through both pigment-pigment and pigment-protein interactions. CD spectra of the mutants in Fig. 4 show that these interactions are maintained despite the blueshift, because the shift in the absorbance peaks is mirrored in the CD spectra. Because CD is sensitive to relatively minor alterations in the geometry of the pigments, we conclude that these mutations have probably not caused any major changes in the conformation or aggregation state of the complexes. Singlet-triplet

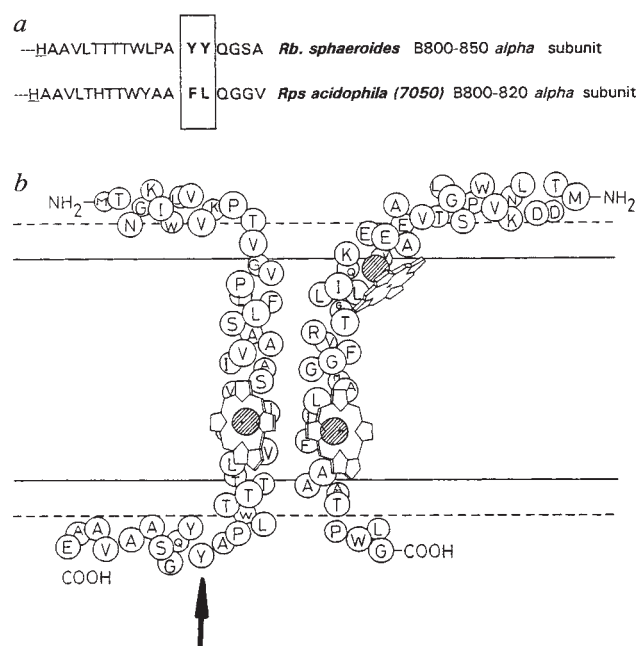


FIG. 1 a, Sequence alignments of the α polypeptides from the LH2 complexes of photosynthetic bacteria suggest that histidine residues bind the bacteriochlorophylls^{7,10}. Near to the binding sites for the two BChl850 molecules are two tyrosine residues at positions 44 and 45 on the α subunit; at the analogous positions for the α subunit of the B800-820 light-harvesting complex of *Rps. acidophila* strain 7050 are the residues Phe and Leu. There are no other consistent differences between the subunits of the B800-820 and B800-850 LH2 complexes, which has led to the suggestion that these tyrosine residues may influence the spectral properties of the bacteriochlorophylls in the LH2 complex⁷. b, Schematic diagram of the LH2 complex of *Rb. sphaeroides* showing the α and β polypeptides, and one B800 (upper) and two B850 (lower) bacteriochlorophylls. The black arrow indicates the position of the α Tyr44 and α Tyr45 residues. As well as the peripheral light-harvesting complex (LH2), which is characterized by two near infrared (NIR) absorbance peaks at 800 and 850 nm due to bacteriochlorophyll (BChl) *a*, *Rb. sphaeroides* also contains a core-light-harvesting complex (LH1) with a single maximum at 880 nm (not shown); LH1 surrounds and interconnects the reaction centres⁸. Both LH1 and LH2 consist of two small polypeptides, designated α and β , of roughly 50 amino-acid residues. Secondary structure prediction methods suggest that α and β both contain a central hydrophobic α -helical stretch that spans the photosynthetic membrane. The low polarization of the fluorescence from the BChls absorbing at 850 nm indicates that the Q_y transitions of these pigments are probably organized as a circularly degenerate oscillator. Linear dichroism studies demonstrate that the Q_x transitions of the BChl850 molecules are roughly perpendicular to the membrane plane⁴.

FIG. 2 77 K absorption spectra of *Rb. sphaeroides* intracytoplasmic membranes prepared from transconjugant strains. Absorbance spectra of membranes from the transconjugant strains where LH2 is the sole pigment protein complex (a) and from transconjugant strains containing the altered LH2 genes in an LH2⁻LH1⁺RC⁺ background (b). It can be seen from the ratio of the 800 to 880 absorption that the altered complexes are produced in a lower ratio to the core complex than the pseudo wild-type LH2, for reasons that are not clear. The gene transfer used has been described previously^{11,12}. The recipient strains used were a, DBC1 (LH2 genes replaced by a kanamycin cassette)¹³, and b, DD13/G1 (LH2 genes replaced by a streptomycin cassette, LH1 and reaction centre genes replaced by a kanamycin cassette)¹⁴. Expression of the altered LH2 genes in the LH2⁻ recipient allows us to evaluate energy transfer from the mutant complex to the native LH1-reaction centre complex, and the spectral properties of the modified complexes can be examined in the other recipient strain free of other interfering pigment-protein complexes. The three LH2 complexes represented are (1) pseudo wild-type (solid line), (2) single-change α Tyr44 $\rightarrow$ Phe mutant LH2 (dashed line) and (3) double-change α Tyr44 $\rightarrow$ Phe, α Tyr45 $\rightarrow$ Leu (dotted line).

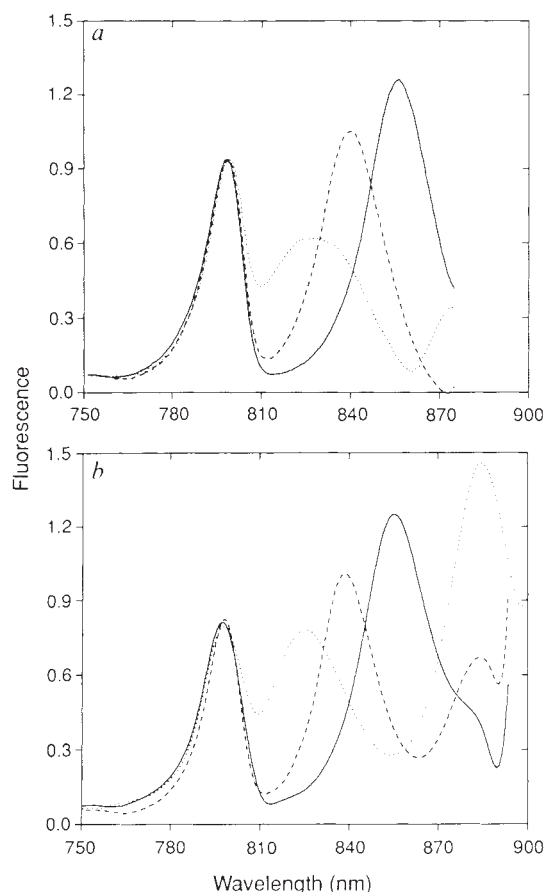
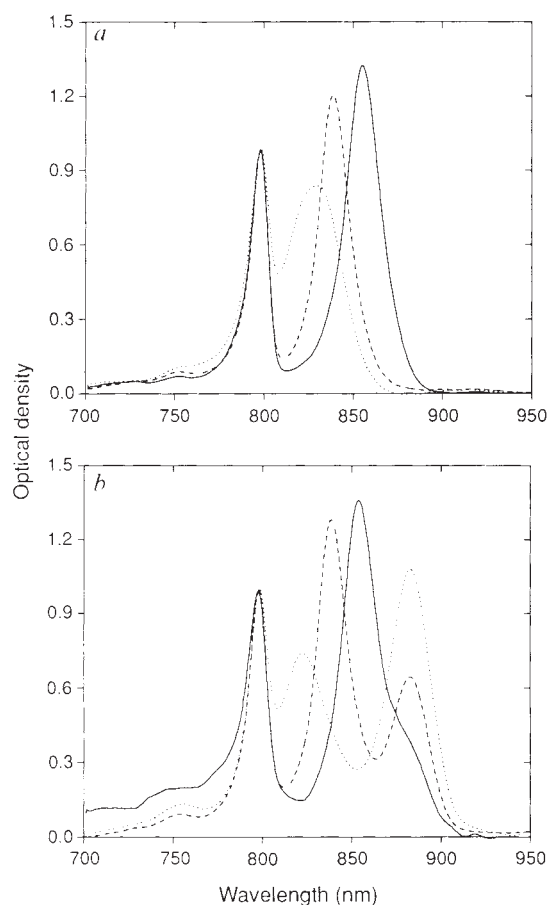


FIG. 3 77 K fluorescence excitation spectra of transconjugant strains: a, strain DBC1; b, DD13/G1 (see legend to Fig. 1). The three LH2 complexes represented are (1) pseudo wild-type (solid line), (2) single-change α Tyr44 $\rightarrow$ Phe mutant LH2 (dashed line) and (3) double-change α Tyr44 $\rightarrow$ Phe, α Tyr45 $\rightarrow$ Leu (dotted line). The spectroscopic apparatus has been described previously¹³. The lower 800/820 ratio in the excitation spectrum compared with the same ratio in the absorbance spectrum is probably due to sample differences, as we have found that this ratio is rather sensitive to sample treatment.

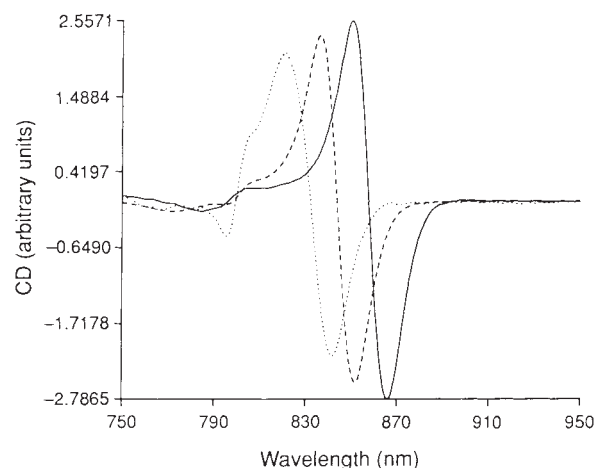


FIG. 4 Circular dichroism spectra of membranes from transconjugant strain DD13/G1 (see legend to Fig. 1). The three LH2 complexes represented are (1) pseudo wild-type (solid line), (2) single-change α Tyr44 $\rightarrow$ Phe mutant LH2 (dashed line) and (3) double-change α Tyr44 $\rightarrow$ Phe, α Tyr45 $\rightarrow$ Leu (dotted line). The spectroscopic apparatus has been described previously¹³. In the double mutant, the signal from the 800 nm bacteriochlorophyll is more intense, suggesting that it is sensitive to the position of the 826 nm peak.

annihilation experiments have shown that both the organization of the bacteriochlorophylls and the domain size of LH2 have not changed because of the mutagenesis (unpublished data).

Brunisholz and Zuber⁷ have drawn attention to the correlation between the occurrence of aromatic residues suggested to be in the vicinity of bacteriochlorophyll *a*, and the redshift of these pigments in light-harvesting complexes *in vivo*. It was suggested that interactions between the π - π^* orbitals of aromatic residues and the tetrapyrrole aromatic rings of the pigments could cause this effect. Calculations on the redshifts on bacteriochlorophylls in the light-harvesting complex of *Prosthecochloris aestuarii* support this suggestion^{8,9}. These results provide direct evidence for

the influence of aromatic residues on the redshift of bacteriochlorophyll *a in vivo*. Modelling studies indicate that α Tyr44 and α Tyr45 lie close to the B850 molecules (Fig. 1); the replacement of these residues leads to a progressively shortened wavelength maximum. The 'new' B800-820 spectrum is still substantially redshifted in relation to monomeric bacteriochlorophyll which absorbs at 770 nm. Thus, the overall red-shifts observed in going from 770 nm (monomeric pigment) to 850 or 880 nm (LH2 or LH1 antenna complex) are probably a function of excitonic coupling between pigments as well as interactions between these pigments and the aromatic residues in the polypeptide backbone. □

Received 24 October; accepted 5 December 1991.

1. Sundström, V. & van Grondelle, R. *J. opt. Soc. Am.* **7**, 1595-1603 (1990).
2. Feher, G., Allen, J. P., Okamura, M. Y. & Rees, D. C. *Nature* **339**, 11-116 (1989).
3. Hunter, C. N., van Grondelle, R. & J. D. Olsen *Trends biochem. Sci.* **14**, 72-76 (1989).
4. Kramer, H. J. M., van Grondelle, R., Hunter, C. N., Westerhuis, W. H. J. & Ames, J. *Biochim. biophys. Acta* **765**, 156-165 (1984).
5. Cogdell, R. G. et al. in *Molecular Biology of Membrane-bound Complexes in Phototrophic Bacteria* (eds Dews, G. & Dawes, E. A.) 211-218 (Plenum, New York, 1990).
6. van Mourik, F. et al. in *Molecular Biology of Membrane-bound complexes in Phototrophic Bacteria* (eds Dews, G. & Dawes, E. A.) 345-356 (Plenum, New York, 1990).
7. Brunisholz, R. A. & Zuber, R. in *Photosynthetic Light-harvesting Systems. Organisation and Function* (eds Scheer, H. & Schneider, S.) 103-114 (De Gruyter, New York, 1988).

8. Pearlstein, R. M. in *Photosynthetic Light-harvesting Systems. Organisation and Function* (eds Scheer, H. & Schneider, S.) 555-566 (De Gruyter, New York, 1988).
9. Gudowska-Nowak, E., Newton, M. D. & Fajer, J. *J. phys. Chem.* **94**, 5795-5801 (1990).
10. Van Grondelle, R., Kramer, H. J. M. & Rijgersberg, C. P. *Biochim. biophys. Acta* **682**, 208-215 (1982).
11. Simon, R., Priefer, U. & Puhler, A. *Biotechnology* **1**, 748-791 (1983).
12. Hunter, C. N. & Turner, G. *J. gen. Microbiol.* **134**, (1988).
13. Burgess, J. G., Ashby, M. K. & Hunter, C. N. *J. gen. Microbiol.* **135**, 1809-1816 (1989).
14. Jones, M. R. et al. *Molec. Microbiol.* (in the press).

ACKNOWLEDGEMENTS. We thank John Olsen for artwork and technical assistance. C.N.H. and G.J.S.F. thank the AFRC for financial support. R.V.G. and R.W.V. thank the Netherlands Organisation for Fundamental Research (NWO) and the Dutch Foundation for Biophysics for their financial support. We also thank the EEC for support.

Selection *in vitro* of single-stranded DNA molecules that fold into specific ligand-binding structures

Andrew D. Ellington & Jack W. Szostak

Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

WE have isolated a set of ligand-binding DNA sequences from a large pool of random sequence DNAs by selection and amplification *in vitro*, using similar methods to those described for the isolation of ligand-binding RNAs¹. The ligand-DNA interactions are both sequence- and ligand-specific, and are dependent on proper folding of the single-stranded DNA. Some ligands led to the isolation of more DNA sequences than RNA sequences, and vice versa. Analysis of individual sequences reveals that ligand binding is DNA-specific; RNAs of identical sequence could not interact with the same ligands. Ligand-binding DNAs might be more suitable than RNAs as potential pharmacological reagents²⁻⁴ because of the greater stability of DNA. The apparent primacy of RNA in the early evolution of life⁵⁻⁷ may have been due to its availability rather than to its functional superiority.

RNA molecules (aptamers) that can specifically bind to a number of small organic dyes have been isolated from a large pool of random sequences¹. This result suggested that a similar approach might lead to the isolation of ligand-specific DNA aptamers, if DNA molecules are also capable of folding into the necessary complex shapes. Repeated cycles of affinity chromatography followed by polymerase chain reaction (PCR) amplification were used to select ligand-binding species from a chemically synthesized pool of random sequence DNA molecules.

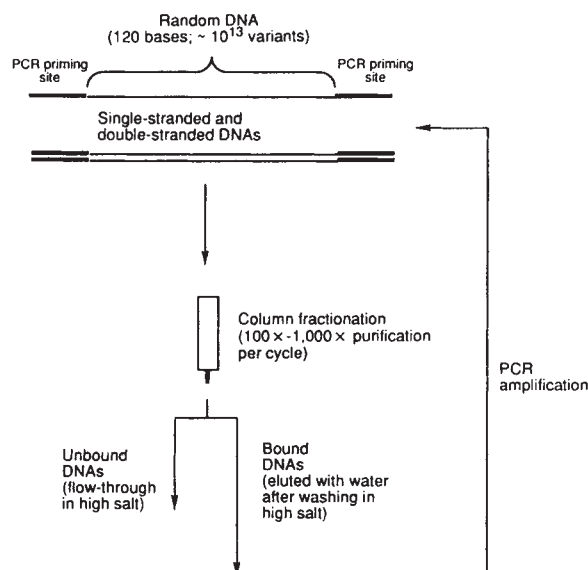
Chemical DNA synthesis of a 157-base oligomer containing 120 bases of random sequence flanked by defined primer-binding sites resulted in an initial pool with an estimated complexity of $2-3 \times 10^{13}$ different sequences (Fig. 1). As we considered that single-stranded DNAs were more likely to fold into complex three-dimensional shapes than double-stranded molecules, we

used an asymmetric PCR amplification⁸ in which replication initiated by one PCR primer was relatively inefficient, leading to the accumulation of single-stranded DNA synthesized from the other primer. To compare the ligand-binding properties of DNA with those of RNA, we used the same affinity resins that we used in the isolation of RNA aptamers: Cibacron blue (CB), reactive green 19 (GR), and reactive blue 4 (B4) coupled to crosslinked agarose beads¹. The large, flat aromatic rings and multiple hydrogen bond donors and acceptors on these ligands provide great scope for productive interactions with nucleic acids.

The DNA pools were applied to affinity columns in high-salt buffer and each column was washed with four volumes of this buffer to remove non-binding DNAs. The retained species were eluted with three volumes of water (to unfold the bound DNAs), precipitated with ethanol and amplified. The DNA pools applied to each column showed increased binding after three to four cycles. But when pools that had been selected for three cycles were applied to non-cognate dye columns to determine whether selected sequences were specifically retained, the pools showed little or no selectivity (Table 1). The majority of DNAs in each pool were apparently nonspecifically bound, either because of binding to the agarose matrix, or because of nonspecific dye binding. We therefore used 'negative selection' to remove nonspecific binding sequences from each pool. Cycle 4 used a pre-column of non-cognate dye suspended over each cognate dye column. The selected pools were first loaded onto the pre-columns and then washed directly onto the cognate dye columns with only one column volume of buffer, so that most nonspecifically bound sequences would remain on the pre-columns. This negative selection significantly enriched the DNA pool in sequences that bound specifically (Table 1). Coupled negative-positive selection was also used in cycle 5 and led to the isolation of pools of DNA that were further enhanced in specificity.

DNA from the final CB and GR selected pools was cloned and individual aptamers were regenerated from plasmid DNAs by PCR amplification using the same primers as during the selection cycles. The aptamers bound with a range of avidities and specificities but no single sequence bound as well as the pool. To eliminate the influence of double-stranded DNAs, native agarose gel electrophoresis was used to separate PCR-amplified single-stranded and double-stranded molecules. The

FIG. 1 Selection scheme. A pool of random DNA molecules (a157.1; 5'-AAGTTCCCGGGCTGCAG(N)₁₂₀TAGGATCCGATTCCTCCC-3') was constructed by chemical synthesis. Short deoxyribo-oligonucleotides were synthesized on a Biosearch 8700 programmable DNA synthesizer (Biosearch) using 1- μ mol columns according to the manufacturer's procedure. Longer fragments such as the 157-base oligomer (157-mer) were synthesized using the same protocol, except that amidites were used at a slightly higher concentration (40 mg ml⁻¹) and the CapA and B reagents were omitted and replaced with CH₃CN in order to carry through deletion events to the end of the synthesis. Based on trityl assays, the overall yield of product was 4%. After deprotection in NH₄OH (55 °C, 16 h) samples were electrophoresed on a 6% acrylamide/7M urea gel and DNA molecules at or near full length were excised. DNA was eluted from the excised band and purified as described¹¹. Yield was 55 μ g (0.1%). A primer extension assay was used to determine the proportion of 157-mer DNA that could be converted into double-stranded DNA and amplified. Excess 157-mer was incubated with labelled primer a36.48 (below) and a modified DNA polymerase (Sequenase; US Biochemical). Only ~5% of the primer could be elongated to full length, as determined using a Betascope 603 Blot Analyzer (Waltham). Increasing the time of deprotection of the 157-mer, or alternative methods of purification, did not affect this result. It seems that the majority of the 157-mer molecules contain chemical lesions that prevent full-length primer extension by DNA polymerase¹². The synthesized DNA was amplified using PCR in a reaction mix of 1.0 μ M primer a36.48: 5'-TAATACGACTCACTA-TAGGAGGAATTCGGATCCTA-3'; 1.0 μ M primer a18.59: 5'-AAGTTCCCG-GGCTGCAG-3'; 400 μ M each dGTP, dATP, dTTP, dCTP; 50 mM Tris-Cl, pH 9.0; 50 mM NaCl; 10 mM MgCl₂ and 5U *Taq* polymerase (Promega). This mix was distributed in 100- μ l aliquots to 200 tubes and cycled through 3 $\times$ (95 °C, 1 min; 45 °C, 2 min; 72 °C, 2 min); and 13 $\times$ (94 °C, 45 s; 55 °C, 1 min; 72 °C, 2 min). Samples were pooled and 1.2 ml (6% of the total) was amplified an additional 5 cycles (55 °C annealing) in the presence of 180 μ Ci of [³²P]-dATP (3,000 Ci mmol⁻¹, Amersham). The radioactively-labelled DNA was isolated on a denaturing polyacrylamide gel. For the first cycle of selection, 3.5 μ g DNA (1.5 pool equivalents) was used for each column; in subsequent cycles, 1–3 μ g DNA was used per column. In each cycle, the selected DNA was



resuspended in 100–600 μ l column buffer (0.5 M LiCl, 20 mM Tris-Cl, pH 7.6, 1 mM MgCl₂), heated at 95 °C for 5 min, allowed to cool to room temperature and then loaded onto columns with bed volumes of roughly 3.5 ml. DNA samples were allowed to equilibrate with the column for 20 min and non-binding sequences were washed away with 13 ml column buffer. Retained sequences were eluted with 10 ml water and precipitated with ethanol. From $\frac{1}{3}$ to $\frac{2}{3}$ of the eluted material from each cycle was used for PCR amplifications; a trial reaction was always run to determine the optimal number of thermal cycles for PCR (the reaction mix and amplification conditions were as described but the number of cycles with 55 °C annealing varied).

TABLE 1 Multiple selection cycles yield specific dye-binding DNA aptamers

	Cycle	CB	GR	Oligo(dT)	B4
	1	2.2	<0.1	0.2	<0.1
	2	<0.1	0.1	0.5	0.1
	3	21.0	<0.1	0.7	0.1
Selectivity without negative selection	→	(2.2)	(1.5)	—	(0.8)
	4	25.0	9.4	32.0	4.5
Selectivity with negative selection	→	(19.0)	(7.0)	—	(4.8)
	5	—	29.0	—	10.0

DNA aptamers were selected on CB, GR and B4 dye-agarose columns and on oligo(dT)_{12–18} cellulose as a positive control. The structures of B4 and CB are given in ref. 6. GR is identical to Procion dye green HE-4BD, the structure of which is given in ref. 7. The per cent DNA bound was calculated by dividing the number of counts eluted from the column in the water wash by the number of counts loaded. Samples were counted using a Betascope 503 Blot Analyzer (Betagen). Cycles 4 and 5 incorporated a pre-selection using 1.2-ml column of a non-cognate dye in series with the cognate dye column. B4 agarose was the 'negative selection' for the CB-selected pool; CB agarose was used for the B4-selected pool; reactive brown 10-agarose was used for cycle 4 and CB-agarose was used for cycle 5 of the GR-selected pool. After loading and equilibration of the sample, the pre-column was rinsed onto the main column with 1.2–1.6 ml of column buffer and the main column was then developed as described in Fig. 1. The input for cycle 5 was taken from the output of the analytical columns (below), effecting an additional purification before cycle 5. The selectivity of the DNA pools is shown in parentheses following cycles three and four, and is defined as (per cent counts bound to cognate dye column)/(per cent counts bound to non-cognate dye column). The same DNA used as input for cycles 4 and 5 (100 μ l in column buffer) was heat denatured, annealed and applied to 1.2 ml cognate and non-cognate dye agarose columns. After 20 min, the columns were rinsed with 3 volumes of column buffer and eluted with 1 volume of 2 mM EDTA, pH 7.6. Each fraction was counted directly in a scintillation counter. The non-cognate dyes used for these tests were: B4 for the selected CB- and GR-pools and CB for the selected B4-pool.

separated DNAs were then assayed for their binding to cognate and non-cognate dye columns (Table 2). The single-stranded species bound efficiently and specifically; the double-stranded DNAs did not. Indeed, the selectivity of the single-stranded molecules for their cognate columns was similar to the selectivity seen previously with RNA aptamers.

We searched for sequence similarities between individual clones to identify motifs responsible for binding. No significant sequence similarities were found by comparison of pairs of 17 CB-binding clones. Variants of an 18-nucleotide (nt) consensus sequence (Fig. 2a) were, however, found in five of eight different GR-binding aptamers. Aptamers containing the 18-nt motif show four to five times more binding than aptamers lacking the motif or departing from the motif (a single base change in GR-32).

We synthesized this 18-nt sequence alone and tested it for binding to the GR resin, but less than 1% was retained on the GR column. Sequence comparison suggested that the 18-nt

TABLE 2 Single-stranded DNA aptamers bind more tightly and with greater specificity than double-stranded DNA aptamers

Aptamer	Per cent DNA bound							
	Single-stranded				Double-stranded			
	CB	GR	B4	BR*	CB	GR	B4	BR*
CB-3	57.0	—	1.0	—	1.1	—	0.3	—
CB-9	27.0	—	0.6	—	1.6	—	0.4	—
CB-106	84.0	—	0.7	—	1.1	—	0.2	—
CB-111	85.0	—	0.9	—	1.5	—	0.2	—
GR-30	—	86.0	0.4	0.3	—	1.0	0.2	0.3

Single-stranded and double-stranded molecules were separated from one another on a 4% NuSieve GTG agarose (American Bioanalytical), 1 $\times$ TBE (0.089 M Tris base, 0.089 M boric acid, 2 mM EDTA) gel. Binding assays are described in the legend to Table 1.

* BR: Reactive brown 10.

a GR-30: 5' GCAGG GGC GTT CGG GGG GTA CCG CTGC 3'
 GR-21: GTACT GTAC
 GR-29: ATTGC CATT
 GR-31: TACCC GGT
 GR-32: CTTAC G TTTG

Consensus sequence: GGC GTT CGG GGG GTA CCG

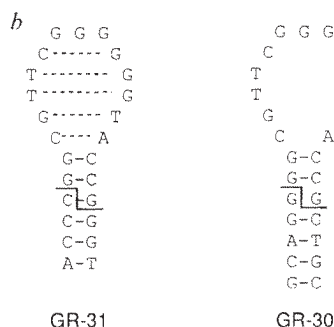


FIG. 2 Sequence motifs in DNA aptamers that bind to GR. a, Several of the GR-binding aptamers contained minor variations of a short 18-base sequence. The variants and their flanking regions are aligned. The bases flanking the consensus sequence are dissimilar but can form paired stems. b, Stem-loops from two of the aptamers were synthesized as 25 nt (a25.30 from the GR-31 aptamer) and 27 nt (a27.32 from the GR-30 aptamer) sequences. These isolated stem-loops bind with high affinity and selectivity to GR-agarose.

sequence formed a stem-loop structure in which the lower portion of the stem is stabilized by the formation of 2–5 additional Watson–Crick base pairs by flanking nucleotides. We therefore synthesized stem-loops that contained the 18-base consensus sequence in a loop closed by 4–5 additional base pairs (Fig. 2). These molecules were found to bind readily to the GR column (>85%). Analytical affinity chromatography⁹ showed that the binding constants (K_d) for the 25- and 27-nt mini-aptamers were $\sim 46 \mu\text{M}$ and $33 \mu\text{M}$, respectively. Binding was sequence-specific in that single-stranded DNA sequences complementary to these aptamers did not bind to the GR resin. In addition, the 18-nt loop appeared to contain most of the specificity determinants as binding was efficient with very different base-paired stems (Fig. 2).

The affinity of DNA molecules for the GR column was a function not only of sequence, but also of chemical composition; less than 1% of RNA transcripts made from either strand of aptamers GR-21, GR-30 or GR-31 bound to a GR column. An RNA version of the 25-nt mini-aptamer also failed to bind to the column. In addition, the sequence motif was not found in our database of previously isolated GR-binding RNA sequences. Similarly, no significant similarities were observed between CB-binding DNA aptamers and CB-binding RNA aptamers. DNA differs from RNA in lacking ribose 2' hydroxyls and in having methyl groups at the 5-position of thymidine. To distinguish which of these could be responsible for the binding difference between identical RNA and DNA sequences, we chemically synthesized a version of the 27-nt mini-aptamer using dU instead of T, which was therefore missing the thymidine methyl group at four positions. This result suggests that it is the presence of 2' hydroxyls that interferes with RNA binding to the ligand, either because of steric exclusion or because the RNA chain adopts a different conformation.

As the initial DNA pool contained $>10^{13}$ random sequences, it should have contained all possible motifs of length 22-nt or less ($4^{22} \approx 10^{13}$). Thus, our results with the GR-binding DNA aptamers suggest that, at least for this ligand, there is a single optimal binding sequence of length 22-nt or less. Of course, there may be longer sequences that can bind more tightly or more specifically but which were not present in the initial pool.

It is of interest to compare the DNA and RNA selections. For one of the ligands (CB) there are apparently a large number of different binding sequences. The minimum number of sequences remaining after selection is the initial pool complexity multiplied by the enrichment factor for each cycle. There would seem to be more DNA than RNA sequences that can bind to CB: roughly 10^{-7} of the original random DNA sequences could bind CB but only 10^{-10} of the random RNA sequences were selected in our previous experiments. In contrast, it seems to be easier to select GR-specific RNA aptamers ($\sim 10^{-8}$ of the original pool) than GR-specific DNA aptamers (10^{-11} – 10^{-12} of the original pool).

Our results indicate that DNA, like RNA, may be able to catalyse chemical transformations. By analogy with catalytic antibodies¹⁰, the isolation of DNA aptamers with sufficient preference for transition-state binding (as opposed to product or substrate binding) could generate new catalysts. □

Received 29 November 1991; accepted 9 January 1992.

1. Ellington, A. D. & Szostak, J. W. *Nature* **346**, 818–822 (1990).
2. Tuerk, C. & Gold, L. *Science* **249**, 505–510 (1990).
3. Abelson, J. *Science* **249**, 488–489 (1990).
4. Riordan, M. L. & Martin, J. C. *Nature* **350**, 442–443 (1991).
5. Gilbert, W. *Nature* **319**, 618 (1986).
6. Visser, C. M. & Kellogg, R. M. *J. molec. Evol.* **11**, 171–187 (1978).
7. Orgel, L. E. & Sulston, J. E., in *Prebiotic and Biochemical Evolution* (eds Kimball, A. P. & Oro, J.), 89–94 (North-Holland, Amsterdam, 1971).
8. Gyllenstein, U. B. & Erlich, H. A. *Proc. natn. Acad. Sci. U.S.A.* **85**, 7652–7656 (1988).
9. Arnold, F. H., Schofield, S. A. & Blanch, H. W. *J. Chromatogr.* **355**, 1–12 (1986).
10. Schultz, P. G., Lerner, R. A. & Benkovic, S. J. *J. chem. Engng News* **68**, 26–40 (1990).
11. Ellington, A. D. in *Current Protocols in Molecular Biology* (eds Ausubel, F. M. et al.) 2.12.1–2.12.5. (Wiley-Interscience, New York, 1987).
12. Green, R., Ellington, A. D., Bartel, D. P. & Szostak, J. W. *Methods* **2**, 75–86 (1991).

ACKNOWLEDGEMENTS. This work was supported by Hoechst AG.

GUIDE TO AUTHORS

Please follow these guidelines so that your manuscript may be handled expeditiously.

Nature is an international journal covering all the sciences. Contributors should therefore bear in mind those readers for whom English is a second language and those who work in other fields. Please write clearly and simply, avoiding unnecessary technical terminology. *Nature's* staff will edit manuscripts to those ends if necessary. Contributors should check their proofs carefully. Because of the competition for space, many of the papers submitted for publication cannot be accepted. For this reason, and because brevity is a great assistance to readers, papers should be as brief as is consistent with intelligibility. Please note that one printed page of *Nature*, without diagrams or other interruptions of the text, has fewer than 1,300 words.

CATEGORIES OF PAPER

Review Articles survey recent developments in a field. Most are commissioned, but suggestions are welcome in the form of a one-page synopsis addressed to the Reviews Coordinator. Length is negotiable in advance but must not exceed six pages of *Nature*.

Articles are research reports whose conclusions are of general interest and which are sufficiently rounded to be a substantial advance in understanding. They should not have more than 3,000 words of text or more than six display items (figures and tables) and should not occupy more than five pages of *Nature*.

Articles should be accompanied by a heading of 50–80 words written to advertise their contents in general terms, to which editors will pay particular attention. A heading (printed in italic type) is not an abstract and should not usually contain numbers or measurements. The study should be introduced in more detail in the first two or three paragraphs, which should also briefly summarize its results and implications.

Articles may contain a few subheadings of two or three words. The meaning of the text should not depend on the subheadings, whose function is to break up the text and to point to what follows. There should be fewer than 50 references.

Letters to *Nature* are short reports of outstanding novel findings whose implications are general and important enough to be of interest to those outside the field. Letters should not have more than 1,000 words of text or more than four display items and should not occupy more than two pages of *Nature*. The first paragraph should describe, in not more than 150 words, the origins and chief conclusions of the study. Letters should not have subheadings or more than 30 references.

Commentary articles deal with issues in, or arising from, research that are also of interest to readers outside the field. Some are commissioned, most are unsolicited. They are normally between one and four pages of *Nature* in length.

News and Views articles are intended to inform non-specialist readers about a recently published advance. Suggestions should be made to the News and Views Editor. Illustrations are welcome. Proposals for meeting reports should be agreed in advance.

Scientific Correspondence is for the discussion of scientific matters, including contributions published in *Nature*. Priority is given to contributions of less than 500 words and five references. Figures are welcome.

Submission. All manuscripts may be submitted to either London or Washington. They should not be addressed to editors by name. Manuscripts or proofs sent by air courier to London should be declared as 'manuscripts' and 'value \$5' to prevent the imposition of import duty and value-added tax (VAT).

GENERAL

All contributions submitted for publication in *Nature* should conform with these rules.

Manuscripts should be typed, double-spaced, with a good-quality printer, on one side of the paper only. Four copies are required, each accompanied by lettered artwork. Four copies of half-tones should be included. Reference lists, figure legends and so on should be on separate sheets, all of which should be double-spaced and numbered. Copies of relevant manuscripts in press or submitted for publication elsewhere should be included, clearly marked as such. Revised and resubmitted manuscripts should also be clearly marked as such and labelled with their reference numbers.

Titles should say what the paper is about with the minimum of technical terminology and in fewer than 80 characters. Authors should avoid active verbs, numerical values, abbreviations and punctuation and should include one or two key words for indexing purposes.

Artwork should be marked individually and clearly with the author's name, figure number and, when known, manuscript reference number. Original artwork should be unlettered. Ideally, no figure should be larger than 28 by 22 cm. Figures with several parts are permitted only if the parts are closely related, either experimentally or logically. Suggestions for cover illustrations, with captions, are welcome. Original artwork (and one copy of the manuscript) will be returned when a manuscript cannot be published.

Colour Artwork. A charge of £500 a page is made for 4-colour figures. Inability to meet these costs will not prevent the publication of essential colour figures if the circumstances are explained.

Figure legends must not exceed 300 words and ideally should be much shorter, and should use telegraphic form wherever possible. The figure should be described first, then, briefly, the method. Reference to a method described elsewhere is preferable to a full description. Methods should not be described in detail in the text.

References should be numbered sequentially as they appear in the text, followed by those in tables and finally those in the figure legends. Reference numbers apply only to papers published or in the press, and a different number should be given to each paper cited. All other forms of reference (including unreferenced abstracts) should be included in the text as personal communication, manuscript in preparation or preprint (with number and institution where appropriate). Text should not be included in the reference list. References should be abbreviated according to the *World List of Scientific Periodicals* (fourth edition, Butterworth, London, 1963–65). The first and last page numbers should be cited. Reference to books should clearly indicate the publisher and the date and place of publication.

Abbreviations, symbols, units and Greek letters should be identified the first time they are used. Acronyms should be avoided as much as possible and, when used, defined. Editors will shorten words if necessary. In case of doubt, SI units should be used.

Footnotes should not be used except for changed addresses or to identify the corresponding author if different from the first-named.

Acknowledgements should be brief. Grant numbers and contribution numbers are not allowed.